

Patrik Floréen
Antonio Krüger
Mirjana Spasojevic (Eds.)

LNCS 6030

Pervasive Computing

8th International Conference, Pervasive 2010
Helsinki, Finland, May 2010
Proceedings



Springer

Lecture Notes in Computer Science

6030

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Alfred Kobsa

University of California, Irvine, CA, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

TU Dortmund University, Germany

Madhu Sudan

Microsoft Research, Cambridge, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max-Planck Institute of Computer Science, Saarbruecken, Germany

Patrik Floréen Antonio Krüger
Mirjana Spasojevic (Eds.)

Pervasive Computing

8th International Conference, Pervasive 2010
Helsinki, Finland, May 17-20, 2010
Proceedings

Volume Editors

Patrik Floréen
University of Helsinki
00014 Helsinki, Finland
E-mail: patrik.floreen@cs.helsinki.fi

Antonio Krüger
German Research Center for Artificial Intelligence (DFKI)
66123 Saarbrücken, Germany
E-mail: krueger@dfki.de

Mirjana Spasojevic
Nokia Research Center
Palo Alto, CA, USA
E-mail: mirjana.spasojevic@nokia.com

Library of Congress Control Number: 2010924981

CR Subject Classification (1998): C.2, H.4, D.2, H.5, C.2.4, H.3

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743
ISBN-10 3-642-12653-7 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-12653-6 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

springer.com

© Springer-Verlag Berlin Heidelberg 2010
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 06/3180

Preface

Welcome to the proceedings of the 8th International Conference on Pervasive Computing (Pervasive 2010). After Toronto, Sydney and Nara, the conference has now returned to Europe. Pervasive is one of the most important conferences in the area of pervasive and ubiquitous computing.

As in the previous year, we had two categories of technical papers: Full Papers and Notes. Pervasive attracted 157 valid submissions, from which the Technical Program Committee (TPC) accepted 24 full papers and one note, resulting in an overall acceptance rate of 16%. The submissions included 628 authors from 27 countries representing all the continents (except Antarctica). As we can see from these figures, Pervasive is a truly global highly competitive conference.

A major conference such as Pervasive requires a rigorous and objective process for selecting papers. This starts with the selection of a high-quality TPC. We were fortunate to be able to draw on the wisdom and experience of our 28 TPC members, from the most prestigious universities and research labs in Europe, North America, and Asia. This committee was aided by the input of no less than 238 external reviewers chosen on the basis of their domain knowledge and relevance to pervasive computing.

The papers were selected using a double-blind review, with four peer reviews per paper, a discussion phase among the reviewers, and a discussion of the papers in the TPC meeting, which was held in Palo Alto during December 12-13, 2009. We thank Nokia Research Center for hosting the meeting.

The keynote of the conference was delivered by Henry Tirri, Senior Vice President, Head of Nokia Research Center. The conference program also included posters, demonstrations and video presentations. The rector of the University of Helsinki gave a reception in connection with the poster, demonstrations and video session. At this event, companies also presented their products.

The first day of the conference was dedicated to workshops and a doctoral colloquium. This year we had an especially rich set of workshops. The last day of the conference included four high-class tutorials.

The conference was organized by the Department of Computer Science at University of Helsinki and Helsinki Institute for Information Technology HIIT, which is a joint research institute of Aalto University and the University of Helsinki. Aalto University is a new university in Finland, resulting from the merger of Helsinki University of Technology TKK, Helsinki School of Economics and the University of Art and Design Helsinki.

We collaborated with the Ubiquitous Computing Cluster Programme in Finland, which is part of the Centre of Expertise Programme (OSKE), and with the ICT SHOK DIEM Programme, which made real-life testing of services for an intelligent conference as part of the arrangements of the conference.

We thank the organizers, sponsors, committee members, reviewers, authors, student volunteers and conference participants for making this all happen.

We especially thank our Conference Manager Greger Lindén for making such an outstanding contribution to the arrangements of the conference.

We hope that the conference provided the participants with an inspiring and interesting program and a pleasant spring-time stay in Helsinki to remember!

May 2010

Patrik Floréen
Antonio Krüger
Mirjana Spasojevic

Organization

Conference Committee

General Chair	Patrik Floréen, HIIT & University of Helsinki
Program Co-chair	Antonio Krüger DFKI GmbH
Tutorials	Mirjana Spasojevic, Nokia Research Center
Demos	Bob Kummerfeld, University of Sydney
Videos	Thomas Strang, German Aerospace Center
Posters	Jonna Häkkilä, Nokia Research Center
Workshops	Tatsuo Nakajima, Waseda University
Doctoral Colloquium	Peter Fröhlich, Telecommunications Research Center Vienna
Proceedings	Gerd Kortuem, Lancaster University
Publicity	Elaine Huang, University of Calgary
Student volunteers	Jukka Riekki, University of Oulu
	Aaron J. Quigley, University of Tasmania
	Petteri Nurmi, HIIT and University of Helsinki
	René Mayrhofer, University of Vienna
	Heikki Ailisto, VTT Oulu
	Andreas Bulling, ETH Zurich
	Alexander De Luca, University of Munich
	Visa Noronen, HIIT
	Vili Lehdonvirta, HIIT
	Florian Alt, University Duisburg-Essen
	Antti Salovaara, HIIT

Program Committee

Louise Barkhuus	University of California San Diego
Nina Bhatti	Hewlett-Packard Labs
A.J. Brush	Microsoft Research
Andreas Butz	University of Munich
Anind Dey	Carnegie Mellon University
Hans Gellersen	Lancaster University
Jonna Häkkilä	Nokia Research Center
Gilian Hayes	University of California Irvine
Elaine Huang	Motorola Labs
Yoshihiro Kawahara	University of Tokyo
Judy Kay	University of Sydney
Julie Kientz	University of Washington
Shin'ichi Konomi	Tokyo Denki University
Marc Langheinrich	University of Lugano (USI)

VIII Organization

René Mayrhofer	Upper Austria University of Applied Sciences
Florian Michahelles	ETH Zurich
Jin Nakazawa	Keio University
Petteri Nurmi	HIIT and University of Helsinki
Patrick Olivier	Newcastle University
Shwetak Patel	University of Washington
Trevor Pering	Intel Research
Albrecht Schmidt	University of Duisburg-Essen
Michael Rohs	Deutsche Telekom Laboratories
James Scott	Microsoft Research
Hide Tokuda	Keio University
Khai Truong	University of Toronto
Alexander Varshavsky	AT&T Labs
Woontack Woo	GIST

Steering Committee

A.J. Brush	Microsoft Research
Hans Gellersen	Lancaster University
Marc Langheinrich	University of Lugano
Anthony LaMarca	Intel Research Seattle
Hide Tokuda	Keio University
Khai Truong	University of Toronto
Aaron Quigley	Human Interface Technology Laboratory

Local Arrangements

Patrik Floréen, General Chair
Andreas Forsblom
Greger Lindén, Conference Manager
Annie Luo, Nokia Research Center, Palo Alto, PC Meeting Arrangements
Petteri Nurmi
Tuija Partonen, Tuhat ja yksi, Registration and General Help
Eeva Peltonen, Tilkutäkki, Web Forms
Jyri Tuulos, Graphic Designer

Reviewers

Gregory Abowd	Stefania Bandini	Fabian Bohnert
Yuvraj Agarwal	Louise Barkhuus	Sebastian Boring
Jasser Al-Kassab	Dominikus Baur	Claus Bossen
Oliver Amit,	Aaron Beach	Anne Braun
Ian Anderson	Michael Beigl	Pamela Briggs
Oliver Baecker	Frank Bentley	AJ Brush
Ronald Baecker	Alastair Beresford	Andreas Bulling
Rafael Ballagas	Nina Bhatti	Andreas Butz

Paul Castro	Jason Hong	Shaun Lawson
Keith Cheverst	Benedikt Hörlner	Lucian Leahu
Tanzeem Choudhury	Gary Hsieh	Dongman Lee
Hao-hua Chu	Elaine Huang	Matthew Lee
Anthony Collins	Pertti Huuskonen	Youngho Lee
Bettina Conradi	Jane Hwang	Jonathan Lester
Sunny Consolvo	Jadwiga Indulska	Kevin Li
Scott Davidoff	Iulia Ion	Stephen Lindsay
Alexander De Luca	Minna Isomursu	Silvia Lindtner
David Dearman	Jhilmil Jain	Linda Little
Mohamed Dekhil	Say Jang	Claire-Michelle Loock
Anind Dey	Matt Jones	Paul Lukowicz
Paul Dourish	Ari Juels	Wayne Lutters
Erica Dubach	Anne Kaikkonen	Lena Mamykina
Nathan Eagle	Apu Kapadia	Jani Mäntyjärvi
Rana el Kaliouby	Amy Karlson	Toshiyuki Masui
Georg Essl	Stephan Karpischek	Peter Mayer
Christian Floerkemeier	Henry Kautz	René Mayrhofer
James Fogarty	Nobuo Kawaguchi	Alexander Meschtscherjakov
Jon Froehlich	Yoshihiro Kawahara	Florian Michahelles
Peter Froehlich	Nao Kawanishi	Dejan Milošević
Kaori Fujinami	Fahim Kawsar	Kazuhiro Minami
Deepak Ganesan	Judy Kay	Masateru Minami
Geri Gay	Joseph Kaye	April Mitchell
Hans Gellersen	Hamed Ketabdar	Keith Mitchell
Sheba George	Julie Kientz	Iqbal Mohomed
Martin Gonzalez-Rodriguez	Sehwan Kim	Jeff Morgan
Paul Grace	Tim Kindberg	Jörg Müller
Saul Greenberg	Evan Kirshenbaum	Tomonori Nagayama
Jonna Häkkilä	Mikkel Kjærgaard	Tatsuo Nakajima
Malcolm Hall	Shinichi Konomi	Jin Nakazawa
Steve Han	Matthias Kranz	David Nguyen
Mat Hans	Sven Kratz	Jeff Nichols
Robert Harle	Christian Kray	Michael Nischt
Gillian Hayes	John Krumm	William Niu
Mike Hazas	Ponnurangam	Petteri Nurmi
Jennifer Healey	Kumaraguru	Kenton O'Hara
Urs Hengartner	Bob Kummerfeld	Patrick Olivier
Jeffrey Hightower	Kai Kunze	Elia Palme
Kenji Hisazumi	Cassim Ladha	Konstantina Papagiannaki
Mark Hodges	Anthony LaMarca	Kurt Partridge
Jesse Hoey	Michael Lamming	Shwetak Patel
Frank Hoffmann	Mik Lamming	Donald Patterson
Paul Holleis	Marc Langheinrich	
	Eric Larson	

Eric Paulos	Shunsuke Saruwatari	Yoshito Tobe
Joao Pedro Sousa	Anthony Savidis	Hideyuki Tokuda
Trevor Pering	Adin Scannell	Zachary Toups
Matthai Philipose	Stuart Schechter	Khai Truong
Andreas Pleuss	John Schettino	Joe Tullio
Thomas Plötz	Bernt Schiele	Ersin Uzun
Erika Poole	Albrecht Schmidt	Kaisa Vaananen-
Christina Pöpper	Johannes Schöning	Vainio-Mattila
Susanna Prittikangas	Eve Schooler	Anna Vallgårda
Daniele Quercia	James Scott	Kristof Van Laerhoven
Aaron Quigley	Masatoshi Sekine	Toni Vanhala
Krishnan Ramanathan	Junaith Shahabdeen	Alexander Varshavsky
Anand Ranganathan	Mike Sharples	Frank Vetere
Madhu Reddy, Carson Reynolds	Naoki Shinohara	Nicolas Villar
Heather Richter	Rich Simpson	Felix von Reischach
Hendrik Richter	Joshua Smith	Stephan von Watzdorf
Till Riedel	Timothy Sohn	Katarzyna Wac
Jukka Riekki	Thad Starner	Juergen Wagner
Jonna Riitta Häkkilä	Thomas Strang	Steven Wall
Jennifer Rode	Jing Su	Jamie Ward
Yvonne Rogers	Anbumani Subramanian	Alexander Wiethoff
Christian Rohner	Shivani Sud	Amanda Williams
Michael Rohs	Yasuyuki Sumi	Woontack Woo
Mario Romero	Petra Sundström	Ken Wood
Barbara Rosario	Rahul Swaminathan	Tatsuya Yamazaki
Virpi Roto	Kaz Takashio	Svetlana Yarosh
George Roussos	Michiharu Takemoto	Keiichi Yasumoto
Enrico Rukzio	Hiroshi Tamura	Koji Yatani
Jung-hee Ryu	Charlotte Tang	Hyoseok Yoon
Alireza Sahami	Tsutomu Terada	Jaeseok Yun
Antti Salovaara	Lucia Terrenghi	Xuemei Zhang
T. Scott Saponas	Georgios Theocharous	
	Niwat Thepvilajanapong	

Best Paper Award Nominees

Jog Falls: A Pervasive Healthcare Platform for Diabetes Management

Lama Nachman, Amit Baxi, Sangeeta Bhattacharya, Vivek Darera, Piyush Deshpande, Nagaraju Kodalapura, Vincent Mageshkumar, Satish Rath, Junaith Shahabdeen, Raviraja Acharya

Virtual Compass: Relative Positioning to Sense Mobile Social Interactions

Nilanjan Banerjee, Sharad Agarwal, Paramvir Bahl, Ranveer Chandra, Alec Wolman, Mark Corner

Common Sense Community: Scaffolding Mobile Sensing and Analysis for Novice Users

Wesley Willett, Paul Aoki, Neil Kumar, Sushmita Subramanian, Allison Woodruff

Table of Contents

Positioning

Virtual Compass: Relative Positioning to Sense Mobile Social Interactions	1
<i>Nilanjan Banerjee, Sharad Agarwal, Paramvir Bahl, Ranveer Chandra, Alec Wolman, and Mark Corner</i>	
The Geography of Taste: Analyzing Cell-Phone Mobility and Social Events	22
<i>Francesco Calabrese, Francisco C. Pereira, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti</i>	
Indoor Positioning Using GPS Revisited	38
<i>Mikkel Baun Kjærgaard, Henrik Blunck, Torben Godsk, Thomas Toftkær, Dan Lund Christensen, and Kaj Grønbæk</i>	

Navigation and Tracking

Specification and Verification of Complex Location Events with Panoramic	57
<i>Evan Welbourne, Magdalena Balazinska, Gaetano Borriello, and James Fogarty</i>	
Tactile Wayfinder: Comparison of Tactile Waypoint Navigation with Commercial Pedestrian Navigation Systems	76
<i>Martin Pielot and Susanne Boll</i>	

Applications

Jog Falls: A Pervasive Healthcare Platform for Diabetes Management...	94
<i>Lama Nachman, Amit Baxi, Sangeeta Bhattacharya, Vivek Darera, Piyush Deshpande, Nagaraju Kodalapura, Vincent Mageshkumar, Satish Rath, Junaith Shahabdeen, and Raviraja Acharya</i>	
EyeCatcher: A Digital Camera for Capturing a Variety of Natural Looking Facial Expressions in Daily Snapshots	112
<i>Koji Tsukada and Maho Oki</i>	
TreasurePhone: Context-Sensitive User Data Protection on Mobile Phones	130
<i>Julian Seifert, Alexander De Luca, Bettina Conradi, and Heinrich Hussmann</i>	

Tools, Modelling

Recruitment Framework for Participatory Sensing Data Collections <i>Sasank Reddy, Deborah Estrin, and Mani Srivastava</i>	138
Out of the Lab and into the Fray: Towards Modeling Emotion in Everyday Life <i>Jennifer Healey, Lama Nachman, Sushmita Subramanian, Junaith Shahabdeen, and Margaret Morris</i>	156
The Secret Life of Machines – Boundary Objects in Maintenance, Repair and Overhaul <i>Matthias Betz</i>	174

Studies

Automatic Assessment of Cognitive Impairment through Electronic Observation of Object Usage <i>Mark R. Hodges, Ned L. Kirsch, Mark W. Newman, and Martha E. Pollack</i>	192
Further into the Wild: Running Worldwide Trials of Mobile Systems ... <i>Donald McMillan, Alistair Morrison, Owain Brown, Malcolm Hall, and Matthew Chalmers</i>	210
Studying the Use and Utility of an Indoor Location Tracking System for Non-experts <i>Shwetak N. Patel, Julie A. Kientz, and Sidhant Gupta</i>	228

Activity Recognition

Object-Based Activity Recognition with Heterogeneous Sensors on Wrist <i>Takuya Maekawa, Yutaka Yanagisawa, Yasue Kishino, Katsuhiko Ishiguro, Koji Kamei, Yasushi Sakurai, and Takeshi Okadome</i>	246
GasSense: Appliance-Level, Single-Point Sensing of Gas Activity in the Home <i>Gabe Cohn, Sidhant Gupta, Jon Froehlich, Eric Larson, and Shwetak N. Patel</i>	265
Transferring Knowledge of Activity Recognition across Sensor Networks <i>T.L.M. van Kasteren, G. Englebienne, and B.J.A. Kröse</i>	283

Sensing

Common Sense Community: Scaffolding Mobile Sensing and Analysis for Novice Users	301
<i>Wesley Willett, Paul Aoki, Neil Kumar, Sushmita Subramanian, and Allison Woodruff</i>	
Active Capacitive Sensing: Exploring a New Wearable Sensing Modality for Activity Recognition	319
<i>Jingyuan Cheng, Oliver Amft, and Paul Lukowicz</i>	
Using Height Sensors for Biometric Identification in Multi-resident Homes	337
<i>Vijay Srinivasan, John Stankovic, and Kamin Whitehouse</i>	

Resource Awareness

Supporting Energy-Efficient Uploading Strategies for Continuous Sensing Applications on Mobile Phones	355
<i>Mirco Musolesi, Mattia Piraccini, Kristof Fodor, Antonio Corradi, and Andrew T. Campbell</i>	
Efficient Resource-Aware Hybrid Configuration of Distributed Pervasive Applications	373
<i>Stephan Schuhmann, Klaus Herrmann, and Kurt Rothermel</i>	

Interaction

12Pixels: Exploring Social Drawing on Mobile Phones	391
<i>Karl D.D. Willis and Ivan Poupyrev</i>	
No-Look Notes: Accessible Eyes-Free Multi-touch Text Entry	409
<i>Matthew N. Bonner, Jeremy T. Brudvik, Gregory D. Abowd, and W. Keith Edwards</i>	
On the Use of Brain Decoded Signals for Online User Adaptive Gesture Recognition Systems	427
<i>Kilian Förster, Andrea Biasiucci, Ricardo Chavarriaga, José del R. Millán, Daniel Roggen, and Gerhard Tröster</i>	

Author Index	445
---------------------------	------------

Virtual Compass: Relative Positioning to Sense Mobile Social Interactions

Nilanjan Banerjee¹, Sharad Agarwal², Paramvir Bahl², Ranveer Chandra²,
Alec Wolman², and Mark Corner³

¹ University of Arkansas Fayetteville

² Microsoft Research Redmond

³ University of Massachusetts Amherst

`nilanb@uark.edu`, `{sagarwal,bahl,ranveer,alecw}@microsoft.com`,
`mcorner@cs.umass.edu`

Abstract. There are endless possibilities for the next generation of mobile social applications that automatically determine your social context. A key element of such applications is ubiquitous and precise sensing of the people you interact with. Existing techniques that rely on deployed infrastructure to determine proximity are limited in availability and accuracy. Virtual Compass is a peer-based relative positioning system that relies solely on the hardware and operating system support available on commodity mobile handhelds. It uses multiple radios to detect nearby mobile devices and places them in a two-dimensional plane. It uses adaptive scanning and out-of-band coordination to explore trade-offs between energy consumption and the latency in detecting movement. We have implemented Virtual Compass on mobile phones and laptops, and we evaluate it using a sample application that senses social interactions between Facebook friends.

1 Introduction

Imagine a suite of social applications running on your mobile phone which senses your precise social context, predicts future context, and logs and recalls social interactions. The possibilities for such applications are myriad [1], from alerting you about an impending contact with a business associate and reminding you of their personal details, to a game that utilizes the relative physical positioning of its players, or a service that tracks the frequency and tenor of interactions among colleagues and friends.

These next generation applications will use continual sensing of social context at an extremely fine granularity. Recent examples of mobile social applications include Loopt [2] which displays the location of a user's friends and Dodgeball [3] which finds friends of friends within a 10 block radius. Unfortunately, these and other widely deployed technologies that implement localization on mobile handhelds are limited by accuracy, coverage and energy consumption.

The most widely used localization technology in mobile handsets is GPS, but it rarely works indoors. Furthermore, its accuracy degrades in urban environments,

and the energy consumed by GPS devices is a significant deterrent. Cell-tower based localization [4] is widely available but can provide very poor accuracy without a fingerprint profile, or outside city centers. Wi-Fi localization, when available, provides reasonable accuracy in dense urban environments, but is also much less effective in other areas [5].

People spend the majority of their time indoors. As a result, many of the most common opportunities for social interaction occur in indoor environments such as offices, hotels, malls, restaurants, music and sports venues, and conferences. In these environments, to detect the interaction with, or even the opportunity to interact with someone requires relatively fine-grained location accuracy. Even in environments where indoor Wi-Fi based localization schemes [6] could provide the needed coverage and accuracy, most of today's environments do not have this infrastructure deployed and the barriers to deployment lead us to believe that this will be the case for some time to come. Techniques that rely on ultrasound or detecting the phase offset of transmitted radio waves [7] are difficult to implement using the hardware and APIs available on commodity mobile phones.

We present the design and implementation of Virtual Compass, a peer-based localization system for mobile phones. Virtual Compass does not require any infrastructure support, but instead uses multiple, common radio technologies to create a *neighbor graph*: a fine grained map of the relative spatial relationships between nearby peers. Virtual Compass allows nearby devices to communicate directly, and provides multi-hop relaying so that the neighbor graph can include others who are not within direct communication range.

Virtual Compass leverages short-range radio technologies, such as Bluetooth and Wi-Fi, available in today's mobile handhelds. These radios consume a significant amount of energy during scanning, and we consider energy management as a fundamental design pillar. Hence, Virtual Compass includes three techniques to reduce energy consumption: 1) use of adaptive scanning triggered on topology changes to update the neighbor graph; 2) selection of the appropriate radio based on its energy consumption characteristics; and 3) using the wide-area wireless network when available with a cloud-based service to assist with coordination and notification of potential changes to the neighbor graph.

Mobile social applications are heavily driven by the *relative* positioning of people, and less by *absolute* location. Sensing the precise placement of individuals relative to one another yields the social context needed for many useful applications, and the quality of location information produced by Virtual Compass increases as the density of devices increases.

We have implemented Virtual Compass on Windows based mobile phones and laptops. Through extensive experimentation we evaluate the latency, location accuracy, and energy consumption characteristics of Virtual Compass as a function of system scale. In a typical experiment we found the average error in spatial placement of nine nodes in a $100m^2$ area was 1.9 meters. We show significant accuracy gains in simultaneously using multiple radios for distance estimation, and our algorithm for spatial placement. Additionally, we are able to locate a new device within 25 seconds of its arrival. Applying our energy conservation

techniques yields four-fold to nine-fold improvements in battery lifetime over a naive scheme that does not use any energy management. We also present the design and evaluation of a sample application built on top of Virtual Compass.

2 Related Work

As a key ingredient for sensing, localization has been the subject of extensive work, both core technologies, and systems that leverage and reason about location. A comprehensive review of localization research is in [8]. Here, we compare and contrast our work by broadly dividing the corpus of prior work into two categories: infrastructure-based and peer-based, and review the most relevant.

Infrastructure-based localization techniques can be broadly classified by their core technology: GSM [9, 10, 5], Wi-Fi [11, 12, 6], GPS, ultrasound with RF [13, 14], Infrared [15], RFID [16], and UWB [17]. The most successful techniques have leveraged infrastructure that was put in place for other reasons (GSM and Wi-Fi localization) and it seems likely that peer-based localization will follow a similar trend relying on technologies such as Wi-Fi and Bluetooth. GPS is the only exception, but it is unique in that it only works outdoors. Several industrial startups [3, 2, 18, 19] have cropped up which use localization to support social applications, relying on the infrastructure-based localization support in mobile phones which is typically Wi-Fi-, GSM- or GPS-based. However, such schemes are limited in coverage and accuracy, making it impossible to support the full range of social applications—especially in situations that require fine-grained proximity information. For example, Wi-Fi localization requires a dense deployment of access points and accurate profiling (not available in many indoor scenarios), and GSM localization can exhibit poor accuracy without a detailed profile or away from dense urban areas.

Peer-based localization techniques attempt to either infer the proximity of a pair of devices, or infer the actual distances between multiple devices and place them in a virtual map. Proximity-based placement schemes such as Hummingbird [20] and NearMe [21] detect if two devices are within 30 to 100 meters of each other. Beep Beep [22] achieves high accuracy using sound, but does not spatially place more than two nodes, nor nodes that are out of earshot. BlueHoo [23] uses Bluetooth discovery to detect friends within Bluetooth range and People-Tones [24] uses GSM-based relative positioning. Virtual Compass measures the distances between multiple nearby nodes, uses multi-hop communication to expand coverage, and spatially places them relative to each other on a 2D plane. Moreover, our system uses algorithms which balance energy consumption with low-latency and accurate localization. Relate [25] and DOLPHIN [26] rely on custom ultrasound hardware which is typically unavailable in commodity devices. RIPS [7] requires signal processing of received radio waves, which is possible on custom hardware such as MICA2 motes but hard to do with off the shelf mobile phones and standard SDKs. MSP [27] uses sensor event distribution to locate nodes in a static sensor network. Bulusus [28], Sextant [29] and Calibree [30] use the location of a subset of nodes (e.g. equipped with GPS units) to derive the

locations of a larger set of nodes. LOCALE [31] also uses GPS equipped nodes to locate other nodes using dead reckoning. Our goal is to design a peer-based localization system that works in the absence of fixed infrastructure or reference points, which can be hard to obtain using GPS in indoor settings.

Moore et al. [32], Spotlight [33], and Vivaldi [34] address the problem of placing nodes relative to each other in a multi-dimensional plane. Moore et al. [32] and Spotlight [33] use custom sensors for relative localization while Virtual Compass focuses on commodity cellular phones. While their algorithms can be used in Virtual Compass, we use a simpler Vivaldi [34] variant in our implementation.

3 Peer Localization

The goal of Virtual Compass is to generate a two-dimensional layout of nearby mobile devices. It uses radios that allow peer-to-peer communication, such as Wi-Fi and Bluetooth, to exchange messages directly between devices. This exchange serves two purposes. Each pair of devices that are in communication range uses the received signal strength of these messages to estimate the distance between them. The message itself contains the list of neighbors and their distances, which allows nodes that are further away to map devices that are not in their immediate communication range. Virtual Compass leverages the collective knowledge of distances between peers learned in this way to calculate the 2D layout.

Figure II shows an example. Mobile node A periodically sends messages to its neighbors B, C, and D. Each of these nodes uses the received signal strength indication (RSSI) of these messages to calculate its distance to A, as described in § 3.1. The nodes exchange these messages on multiple radios to reduce the inherent error of distance estimation via RSSI, as described in § 3.2. They embed these distances in the messages that are exchanged between neighbors so that each node discovers the distances between other nodes. So in this way, C learns of the distance between A and D. Furthermore, nodes such as E that are far away can learn where A, B and D are. Virtual Compass solves the constraints imposed by these distances to create a relative map using the technique in § 3.3. Note that the underlying RSSI-based mechanism detects distance but not direction.

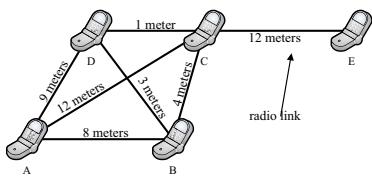


Fig. 1. Each line represents a mobile node’s ability to directly communicate between the two end-points using a radio such as Bluetooth or Wi-Fi. A,B,C and D are in communication range of each other, while E is only in range of C

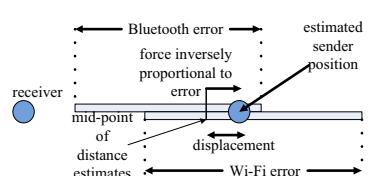


Fig. 2. An example of using RSSI measurements from multiple radios (Bluetooth and Wi-Fi) to reduce the error in computing distance

3.1 Estimating Distance

In Virtual Compass, nodes periodically exchange messages on radios with omnidirectional antennas. § 5 describes these messages in detail. We use the RSSI of these messages to estimate the distance between sender and receiver. Even though we rely on RSSI, if techniques such as *propagation time* become feasible, Virtual Compass can easily use them instead. While translating RSSI to distance has been studied in prior work [35, 36], Virtual Compass enhances that work by incorporating the uncertainty in distance measurements to provide two benefits. First, as Virtual Compass is meant to be used in a broad range of unknown environments, modeling the uncertainty reduces the dependence on the environment in which the measurements were taken. Secondly, and more importantly, the error model provides a basis for composing information from different radios. To translate each RSSI reading to a distance estimate with an error bound, we use empirical models that we built by running several propagation experiments in two indoor office environments at Microsoft Research Redmond, and University of Massachusetts Amherst (details of the experiments can be found in our technical report [37]). We have evaluated our distance estimation scheme in § 6.

3.2 Using Multiple Radios

To reduce the error in estimating distance from RSSI, Virtual Compass uses multiple peer-to-peer radios simultaneously. For ease of exposition, we describe how our scheme works for two radios, Wi-Fi and Bluetooth. This approach works for any radio with an RSSI to distance conversion, or when using more than two radios.

Consider Figure 2 where a node receives a message from the sender over Bluetooth and one over Wi-Fi and attempts to calculate the distance between the two nodes. Let $RSSI_1$ be the RSSI of the message received over Bluetooth, and $RSSI_2$ be the RSSI of the message received over Wi-Fi. We obtain a distance estimate for each, x_1 and x_2 (see [37] for details). We also obtain the uncertainties (error), u_1 and u_2 , each of which is the distance between the 10th and 90th percentiles for the measured model. The goal of the composition is to combine the two sources of information in order to reduce uncertainty in measurement. The mid-point of the two RSSI distance estimates is $P = (x_1 + x_2)/2$. We apply a displacement from P for each measurement, which are $F_1 = (P - x_1) * u_1/2$ and $F_2 = (P - x_2) * u_2/2$. Intuitively, the sum of the forces should push the node in the direction of a source which has a smaller uncertainty in measurement. The final estimate of the distance is given by the midpoint displaced by a normalized sum of displacements $D = P + 2(F_1 + F_2)/(u_1 + u_2)$. The normalization ensures that the estimate of distance always falls within the range of estimates given by the two RSSI readings. In the rare case where the uncertainties from the two readings do not intersect, we simply use P as the final distance. We have evaluated our multi-radio composition scheme in § 6.

In this way, each pair of nodes that can directly communicate with each other estimate the distances between them, while reducing error. These distances are

embedded in the messages that are exchanged between them so that ultimately, each node knows the distances between any two nodes that can communicate in the vicinity. The next step in Virtual Compass is to calculate a 2D spatial placement of these nodes that satisfies these distance constraints.

3.3 Spatial Placement

Consider a 2D Euclidean space where each node's position is determined by its (x, y) coordinates. Each distance estimate r_{ij} between nodes i and j forms a constraint: $(x_i - x_j)^2 + (y_i - y_j)^2 = r_{ij}^2$. An optimal algorithm would simultaneously solve this set of *non-linear* (quadratic) constraints to calculate coordinates for peer nodes. However, this is known to be NP-Hard [38]. Furthermore, since the distances between nodes are measured independently and are subject to error, it is possible in some cases that there is no solution that satisfies all the distance constraints.

We instead use the Vivaldi [34] method to calculate node positions. Vivaldi uses estimates of distances between nodes to calculate a force vector, and then iteratively improves each node's coordinates by moving it along the resulting force. Vivaldi has been shown to produce good results with little computation overhead. However, the choice of starting all nodes at the origin can sometimes lead to local minima or a large number of iterations to converge. Hence, to produce a relative map of all nodes, we first calculate a very approximate but quick placement in *phase 1*, and then feed that to a simple Vivaldi implementation in *phase 2* for iterative refinement.

Phase 1 calculates an approximate set of coordinates that will help Vivaldi converge faster and to more accurate results in phase 2. Consider the example where node A is calculating a placement for itself with respect to 2 other nodes B and C and begins by placing itself at the origin. It finds the peer, *B*, that is the smallest distance (r_1) away, and places it at $(0, r_1)$. Next, we choose node *C* that is constrained by both *A* and *B*. The algorithm Virtual Compass uses to place *C* is defined in Algorithm II. We run this algorithm multiple times with different constraint orderings and we use an average of the coordinates from each

Algorithm 1. Spatial placement for calculating rough 2D coordinates during Phase 1 for a node

```

Input: Set of constraints  $C = \{C_1, \dots, C_n\}$ ,  $C_i = (x - x_i)^2 + (y - y_i)^2 = r_i^2$ 
loop
    For every pair of constraints  $(C_i, C_j)$ , find intersection points  $(x_1, y_1)$  and  $(x_2, y_2)$ 
end loop
 $P = \{\{(x_1^1, y_1^1), (x_2^1, y_2^1)\}, \dots, \{(x_1^k, y_1^k), (x_2^k, y_2^k)\}\}$  (set of intersection coordinates).
Initialize solution set  $S = \{(x_1^1, y_1^1)\}$ 
loop
    For each element  $E = \{(x_1^j, y_1^j), (x_2^j, y_2^j)\} \in P$ 
         $S = S \cup \arg \min \left\{ \sum_{(x_j, y_j) \in S} \sqrt{(x_1^j - x_j)^2 + (y_1^j - y_j)^2}, \sum_{(x_j, y_j) \in S} \sqrt{(x_2^j - x_j)^2 + (y_2^j - y_j)^2} \right\}$ 
end loop
return Node coordinate:  $((1/|S|) \cdot \sum_{x_i \in S} x_i, (1/|S|) \cdot \sum_{y_i \in S} y_i)$ 

```

iteration as the starting placement for phase 2. Experimentally, we determined that 10 iterations produces a sufficiently accurate initial placement with little impact on run time. While we could have used other algorithms, the goal of this phase is to produce a starting point for Vivaldi that is more reasonable than the origin for all nodes.

Phase 2 uses the coordinates from phase 1 as the starting placement and uses a simple implementation of Vivaldi [34] to iteratively refine the coordinates to reduce the error between the placement and the measured pairwise distances. In each iteration, Vivaldi calculates forces that are applied between nodes – each force represents the difference between the measured distance between a pair of nodes and their distance in the virtual coordinate space. The resulting force on each node then determines the direction and amount of movement for the node in the virtual coordinate space. This process is repeated in each iteration. We have experimentally determined that 100 iterations produces accurate results with extremely marginal benefit from additional iterations. In § 6, we present the latency overhead of this computation, and it is dwarfed by the network communication time.

As an example, consider node A at (x_1, y_1) with a neighbor B whose coordinates are (x_2, y_2) . The measured distance between them is r_{12} . The magnitude of the force F between them as applied on A is $r_{12} - \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ and its direction is given by the unit vector $((x_1 - x_2), (y_1 - y_2))$. There may be other forces applied on A (due to measured distances to other neighbors), and we calculate the resulting single force [34]. The coordinates for A are then changed in this iteration to $(x_1 + F_x * t, y_1 + F_y * t)$, where F_x and F_y are the components of F in the x and y direction and t is a constant. For Virtual Compass we experimented with different values of t and found $t = 0.1$ works best in our environment. Applying a force at each node that is proportional to the error minimizes the mean-square error and converges to a set of coordinates which satisfy the distance constraints (see [34] for proof).

4 Energy-Efficient Peer Localization

As with any system targeted at mobile devices, energy consumption is a critical concern. If the lifetime of the device is severely impacted, users will eschew applications that rely on our system. Virtual Compass depends on frequent communication between peers to provide timely updates to changes in the social graph.

Table 1. Energy consumption of radios on fully-charged HTC Touch phones

Radio	Power (mW)	Lifetime (hours)
GSM (idle)	24.4	203.0
Bluetooth (idle)	45.5	109.8
Bluetooth (scanning)	507.6	9.8
Wi-Fi (idle)	849.9	5.9
Wi-Fi (scanning)	1305.4	3.8
GPS (idle)	859.9	5.8
GPRS (transfers)	1031.2	4.8
HSDPA (transfers)	1099.6	4.5

As has been observed in prior work [30], communication consumes a significant portion of a mobile phone’s energy budget. To place our work in a common frame of reference, we include Table II which shows the energy consumption of our implementation platform. With no communication, a typical phone will last for 203 hours on a single battery charge. However, if it continuously scans for other peer devices, the battery is completely exhausted within 10 hours when using Bluetooth, and under 4 hours using Wi-Fi.

To mitigate this, Virtual Compass must balance the energy devoted to sensing and maintaining the social graph against the accuracy of the system. Scans that are too frequent will drain energy, and scans that are too infrequent will increase the latency for peer localization – device arrival or departure will go undetected until the next scan interval. Virtual Compass uses three techniques to reduce the number of scans without significantly degrading localization accuracy.

4.1 Adaptive Bluetooth Scanning

We observe that repeated scans are unnecessary in a static environment, such as when there are no other devices around, or when none are moving. Virtual Compass uses this observation to adapt the scan interval. Every device keeps track of changes in its neighbor graph and accordingly adjusts its scan interval – aggressively scanning the environment when the neighbor graph changes, and increasing the scan interval otherwise. To track the change in its neighbor graph, a device calculates the number of one hop, N_1 (2 paths in Figure B), and two hop paths, N_2 (1 path in Figure B), that have changed between successive scans. We compute a change metric as $p * N_1 + (1 - p) * N_2$, where p is a constant. When this metric is less than a threshold x , we increase the inter-scan interval by 10 seconds. If the metric is above a threshold y , we halve the scan interval. We do not scan more frequently than once every 10 seconds and we do not allow the scan interval to increase beyond 10 minutes. We use values of 0.9, 1 and 1 for p , x and y respectively. While these values are arbitrary and could be tied to an application, they work well in our experiments. The results of a simple experiment showing the behavior of this technique are shown in Figure 4. The scan interval additively increases until new devices are introduced or removed in the neighbor graph, at which point the scan interval is halved. This method can be easily extended to other metrics. For instance, an application may care about sensing small changes in the distance between peers or may want to weight different peers based on their significance in the social network.

Between two successive scans, which can be as long as 10 minutes, we leave the Bluetooth radio on since the idle energy consumption of Bluetooth is small (see Table II). Moreover, with the Bluetooth radio on, it can respond to its peer’s scans and the corresponding neighborhood graph is always complete. In contrast, the idle power consumption for Wi-Fi is comparable to scanning. Therefore, the radio needs to be turned off between scans. However, this implies that adaptive scanning for Wi-Fi is infeasible—if different peers wake up at different times, their scans will result in incomplete neighbor graphs. Therefore, for Wi-Fi we periodically (every 1 minute at wall clock time) turn on the radio and put it

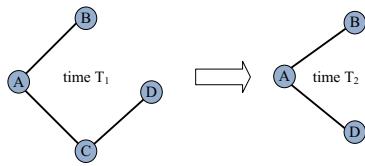


Fig. 3. An example illustrating the calculation of the change metric for adapting the Bluetooth scan interval. In this example, the change metric between T_1 and T_2 is 1.9, which is high enough to trigger a reduction in the scan interval.

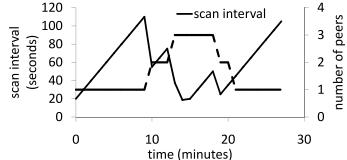


Fig. 4. This graph shows the Bluetooth scan interval for a device over time. After 10 minutes, we introduced a second peer device, and at 12 minutes a third peer device. At 18 minutes we removed one device, and then a second one at 20 minutes.

in scan mode. Mobile phones synchronize their wall clock time with the cellular infrastructure. Even if disconnected from the cellular network, clock drift in the order of one or two seconds is not a significant issue since Wi-Fi scanning takes several seconds (see § 6).

4.2 Cloud Coordination

There are significant periods of time when a device is completely alone. Figure 5 shows how often Bluetooth scans by 150 participants [40] found other devices. On average, each mobile phone found no other Bluetooth devices 41% of the time. While it is possible that other devices were present but did not have Bluetooth discovery enabled, or were discoverable over the longer range of Wi-Fi, this finding fuels our belief that there are periods of time when a device is completely alone. Hence we can save energy on devices during these periods by keeping Wi-Fi off and not initiating Bluetooth scans until a new device arrives. However, the primary challenge is to detect device arrival without using Bluetooth and Wi-Fi. We observe that many mobile devices are almost always connected to the Internet via a cellular data connection such as 3G. Hence, a simple service running on the Internet can inform the device when there are other devices in the vicinity.

In Virtual Compass, each mobile device uploads its approximate geographic location to this service. This location is calculated using low-energy, coarse grained GSM localization. The list of GSM cellular towers that are in the vicinity and the RSSI values are used to compute a rough geographic location. Each time its location changes, the device updates the service. When the device believes it is alone (no neighbors in Bluetooth and Wi-Fi scans), it will periodically ask this Web service whether there are any other devices in the vicinity running Virtual Compass. If there are no peers around it, the device will keep its Wi-Fi radio off and not scan on Bluetooth. Otherwise, it adjusts its scan interval and Wi-Fi wakeup interval as described previously. Since periodic polling on a radio such as 3G consumes a considerable amount of energy, Virtual Compass uses a push-based technique to notify the device when other nodes are around. Inspired by

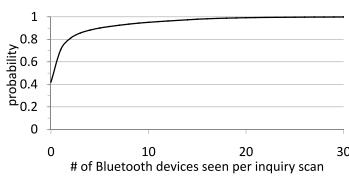


Fig. 5. CDF of the number of Bluetooth devices seen in periodic scans from 150 Nokia N95s [40]

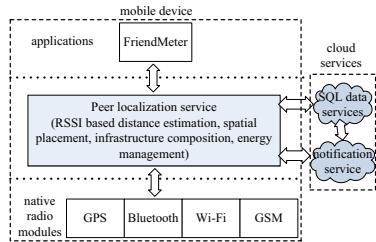


Fig. 6. Architecture of Virtual Compass

Cell2Notify [41], a Virtual Compass device uploads a Request-for-Notification (RFN) bit to the Web service when it thinks it is alone. For each device with the RFN bit set, the Web service keeps track of other device arrivals in the vicinity of the sleeping device and will notify it, which then resumes Wi-Fi and Bluetooth scanning. We describe our implementation of this notification in the next section.

4.3 Leveraging Application Behavior

In addition to exploiting user mobility to reduce energy consumption, a cloud service allows us to also exploit application behavior. Some applications that use peer localization may not need the neighbor graph maintained all the time, even though the applications are still running. For example, an application that shows the user a map of nearby friends and how to get to them does not need the neighbor graph if the user is not interacting with the phone. Scanning in this scenario wastes energy. However, not scanning, and hence not participating in multi-hop discovery, could degrade localization accuracy for other devices where their users are actively interacting with the phone. We suspect that there are significant periods of time when every phone in the vicinity is simultaneously not in use. To detect this scenario, Virtual Compass detects when the back-light for the screen on a mobile device turns off. We then assume that the user is not using the application and upload this bit of information to the Web service along with the device's rough geographic location. When Virtual Compass polls the Web service to find out how many devices are in the vicinity or uses the notification service, it also learns how many of them have the back-light on. If no devices are actively being used, then it keeps the Wi-Fi radio off and does not scan on Bluetooth. If *any one* device in the vicinity has the back-light on, then it resumes normal discovery behavior. Unfortunately, if an application uses peer localization to log social interaction in the background, instead of displaying an interactive map, then this technique cannot be used.

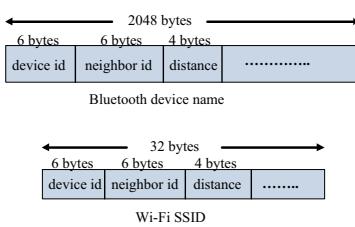


Fig. 7. Packet format for beacons in Virtual Compass



Fig. 8. Screen shots of the simple UI in FriendMeter

5 Implementation

We have implemented Virtual Compass on the Windows Mobile 6.0 operating system that runs on a variety of mobile phones. While we have also ported Virtual Compass to the Windows Vista operating system, we focus on the mobile phone version in this section. Virtual Compass runs entirely at the application layer, and does not require modifications to the Bluetooth and Wi-Fi drivers, nor to the network stack. Our software architecture, depicted in Figure 6, consists of four main components: native radio modules, cloud services, peer localization service and applications.

Native radio modules: Virtual Compass interacts with many radios (GPS, Bluetooth, Wi-Fi, and GSM) using native APIs exposed by the Windows Mobile OS. To access the Wi-Fi radio when the device is in suspension state S3, we use a PPN_UNATTENDED state. This consumes slightly more energy than S3, but allows us to access the Wi-Fi radio.

For device discovery and propagating the neighbor graph, as described in § 3, Virtual Compass requires every device to periodically broadcast its ID and the IDs of and distance to each of its peers. The application layer provides the ID to be used in Virtual Compass. To broadcast this information without the additional latency of explicitly forming a network, we use the Beacon-Stuffing approach [42] for Wi-Fi, and a similar technique for Bluetooth. Our beacon formats for Bluetooth and Wi-Fi are in Figure 7. For Bluetooth beacons we modify the 2048 bytes available for the device name, while for Wi-Fi beacons we embed this information in the 32 byte SSID. The small size of the Wi-Fi SSID limits the size of the neighbor graph that can be encoded in the beacon. To solve this problem, we could use two techniques proposed by Beacon-Stuffing [42]: use 256 byte IE blobs, or fragment large strings across beacons that are then reassembled at the receiver. We have not implemented either solution, and in our current implementation, we limit the neighbor graph embedded in beacons to immediate one-hop peers, thus effectively limiting peer localization to a maximum of two-hops.

One problem with using the Bluetooth radio for peer localization is that it may interfere with Bluetooth headset usage during phone conversations. A scan in the middle of a conversation will disrupt the phone call. To avoid this problem, we

trap the *incoming phone call* and *phone call talking* events from the Windows Mobile OS and stop Bluetooth scanning if either event is active. We resume scanning once these events have ended.

Cloud services: Virtual Compass uses the *SQL Server Database Service* (SSDS)^[43] over the Internet for coordinating Wi-Fi radio wake-ups and Bluetooth scans, as described in § 4.2. SSDS has the following components: (a) Authority: this is the top-most level of containment hierarchy under which all the data for a particular SSDS login is stored. (b) Container: an authority is a collection of containers. (c) Entities: each entity inside a container stores any number of user-defined properties and values. Virtual Compass uses a single authority, under which there is a separate container for each geographic region, under which there is a separate entity for each device. The peer localization service moves the device’s entity to the appropriate container based on cellular tower IDs and RSSIs from the GSM radio and updates a bit indicating whether the screen back-light is on. Virtual Compass can use a push (notification-based) and polling scheme to download information on neighbor positions. For polling it periodically downloads the contents of the containers to determine if it is alone.

When using the notification scheme, each Virtual Compass device uploads its current position based on cell tower IDs and RSSIs. When a device does not find any neighbors on a Bluetooth and Wi-Fi scan it uploads a RFN (*Request for Notification*) bit and the device’s phone number to the cloud and stops scanning on Bluetooth and switches off Wi-Fi. A notification service runs on an Internet server which constantly downloads the location of all Virtual Compass devices using SSDS. It calculates whether any Virtual Compass device is near a node with its RFN bit set. If so, it uses a Skype client on the server to make a phone call to the device using a special caller ID number. The device traps the incoming phone call event, and if it recognizes the special caller ID number, it ends the call and resumes scanning on Bluetooth and Wi-Fi.

Peer localization service: The location service runs the distance estimation and spatial placement algorithms from § 3 to produce a 2D map of where peer devices are. The distance estimation model that we use to convert a RSSI measurement to distance and uncertainty is described in our technical report [37]. We used extensive measurements in two office environments at Microsoft Research Redmond, and University of Massachusetts Amherst to derive these models. The service also manages the Wi-Fi radio sleep and scan schedule, Bluetooth scanning interval and interfaces with the cloud services to reduce energy consumption as described in § 4. It feeds the entire map to the application layer.

Applications: We have implemented the *FriendMeter* application using Virtual Compass. FriendMeter uses Virtual Compass to sense the distances between the user and her friends who are in the vicinity. Several applications such as *gaming* and *file transfer* amongst friends can be considered as instances of FriendMeter. FriendMeter is designed with two purposes in mind – a short-term use and a long-term use for the sensed information. In the short-term, the results from Virtual Compass are used to show the user a map that can be used to find and meet her friends. In the long-term, the time-varying distances measured

between the user and her friends can be used to infer social interactions. These inferences can be used to cluster friends in social applications, such as Facebook, based on proximity. Each friend can be metered by the amount of physical social interaction. Our implementation shows the user a map and records a history of the map, but currently does not alter their friends list.

FriendMeter uses the Facebook API to connect to Facebook, authenticate the user and get her list of friends. It uses a unique numerical Facebook login id—provided by Facebook as the mobile device’s ID. This facilitates identifying the user on each peer device, but as we note in § 7, there are some privacy implications. FriendMeter displays a map of all the user’s friends in the vicinity. It also displays the photographs of the nearby users and their interests, hobbies, and other information. Screen shots from the application are in Figure 8. Even though the underlying peer localization service provides a map with many devices, FriendMeter filters out those that are not in the user’s friend list.

6 System Evaluation

We evaluate the performance of Virtual Compass by focusing on the following three key questions: (1) How accurate are Virtual Compass’s distance estimates and spatial placement? (2) How much energy does Virtual Compass consume? (3) How quickly does Virtual Compass adapt to changes (e.g., when a new device arrives, or one departs)? In answering these questions, we also examine the impact of scale: how does the number of devices affect Virtual Compass?

Experimental Setup: We evaluate Virtual Compass on the Windows Mobile and Windows Vista operating systems. Our testbed consists of ten devices – an HTC TyTNII mobile phone, an HTC Touch Cruise mobile phone, four laptops, and four desktops. All ten devices have IEEE 802.11b and Bluetooth interfaces, and are connected to the Internet via 3G cellular on the phones or Ethernet on the laptops and desktops. In most experiments, we deploy the devices in a $100m^2$ indoor office area, but we also evaluate larger areas of $900m^2$ and $2500m^2$ where indicated. Many experiments involve statically-placed nodes, but in those evaluating latency, we move a device into or out of the deployment area. When evaluating energy consumption, we measure the lifetime of the fully charged mobile phones while running Virtual Compass and leaving the GSM radio on.

Accuracy of Localization: The primary goal of Virtual Compass is to accurately localize nearby peers. We evaluate this accuracy in two ways – (1) *error in pairwise distance* between nodes – what is the difference between the physical distance and the distance that Virtual Compass predicts? (2) *spatial placement*: for a number of nodes, how different is the 2D placement that Virtual Compass presents from their actual placement?

Pairwise distance accuracy: Figure 9 shows how well Virtual Compass estimates the distance between two nodes as their physical distance is varied. Virtual Compass comes very close to perfectly estimating distance. When Virtual Compass does deviate from the actual distance, it does so by a small amount as the

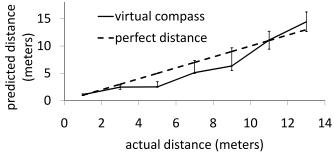


Fig. 9. Distance and deviation predicted by Virtual Compass

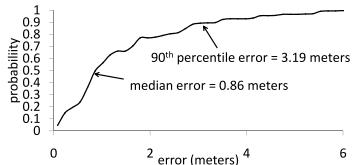


Fig. 10. CDF of the error in distance

Table 2. Average error for nine devices in a $100m^2$ indoor area reported by the different systems

System	Average Error (meters)
Bluetooth	3.40
Wi-Fi	3.91
Virtual Compass	1.41

error bars indicate. Figure 10 shows the CDF of this error over a large number of placements. The median error is only 0.9 meters, and over 90% of the time, the error is under 2.7 meters. To examine why Virtual Compass is so accurate in pairwise distance estimation, we present Table 2, which shows the advantage of our multi-radio approach. If Virtual Compass were to use only Bluetooth radios, the average error would be quite high at 3.40 meters, or 3.91 with just Wi-Fi radios. However, by simultaneously using both Bluetooth and Wi-Fi, Virtual Compass reduces the average error to 1.41 meters.

Spatial placement accuracy: We evaluate spatial placement in Figure 11. Virtual Compass's 2D spatial placement (dark dots) almost exactly matches the actual placement (light dots) – the average distance between a light dot and the

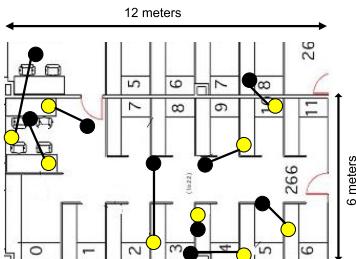


Fig. 11. This figure shows a map of 9 devices using light dots. We overlay the spatial placement map from Virtual Compass on this figure using dark dots. The spatial placement was generated at the node on the desk between “3” and “4”.

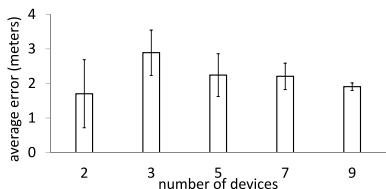


Fig. 12. Using multiple experiments similar to Figure 11, we calculate the average error in 2D placement as we vary the number of devices

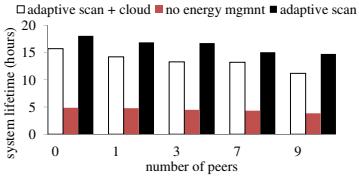


Fig. 13. This graph shows the lifetime of Virtual Compass on a fully-charged stationary mobile phone, with different energy conservation techniques and different numbers of nearby peer devices in a $100m^2$ area. In this experiment, the devices did not move

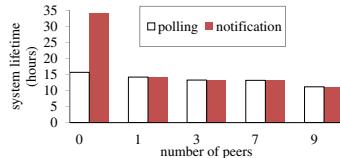


Fig. 14. The lifetime of Virtual Compass on a fully-charged mobile phone for the same setup as Figure 13. In the “polling” bars, Virtual Compass periodically queries the cloud service for node arrivals and in the “notification” bars, Virtual Compass uses the push-based technique.

corresponding dark dot is 1.9 meters. Our accuracy is dependent on two factors: our multi-radio RSSI-based distance estimation, and our 2D spatial placement algorithm. To tease apart these two factors, we applied our 2D spatial placement to the *actual* pairwise distances between these nodes (as opposed to the RSSI-based estimates) and the average error is 0.6 meters.

The accuracy of our 2D placement algorithm also depends on the density of devices – the more devices we have, the more constraints we have that allow the placement to converge faster. In Figure 12, we repeat our placement experiments while varying the number of devices, and the placement of these devices. As the number of devices is lowered, the error increases because every node is constrained by fewer neighbors. With just 2 nodes (with one placed at the origin), the average error is purely a reflection of the RSSI-based distance estimation error.

Energy Consumption: While accuracy in localization is the primary goal of Virtual Compass, energy consumption is a critical concern for mobile devices. We now evaluate the benefits of the energy saving techniques from § 4. Figure 13 shows the lifetime as the number of nearby peers is varied. The “no energy mgmnt” bars, the lifetime of a mobile phone with Wi-Fi and Bluetooth always on and scanning every 1 minute and 10 seconds respectively is dismal, at 4.8 hours with no peers and 3.8 hours with 9 peers. The slight drop in lifetime with the number of peers is because Virtual Compass has to connect over Bluetooth to every peer to get the RSSI value (this is a limitation of the Windows Mobile Bluetooth API). However, when we turn Wi-Fi on and off every 1 minute and adaptively change the Bluetooth scan interval, we see significant energy savings in the “adaptive scan” bars, from 18.0 hours with no peers to 14.8 hours with 9 peers. Even though the devices do not move, there is a drop in lifetime with the number of peers because of the Bluetooth connect issue and because with more devices, variations in the environment can temporarily appear as slight neighbor graph changes. When we include the cloud coordination scheme in the “adaptive scan + cloud” bars, the lifetime actually reduces. Periodically

Table 3. This figure shows the lifetime of Virtual Compass on a fully-charged mobile phone, with the back-light optimization from § 4.3 turned off or on. We used a synthetic workload based on the Reality Mining data [44] to emulate phone usage

back-light optimization	lifetime (hours)
off	12.07
on	15.42

Table 4. This Figure shows the lifetime of Virtual Compass on a fully-charged mobile phone, with 9 peers nearby, across different sizes of regions. In each experiment, the devices did not move.

Density (meter ⁻²)	Lifetime (hours)	1-hop peers	2-hop peers	3-hop peers
100	11.19	9	0	0
900	11.92	5	4	0
2500	12.05	5	3	1

polling the Web service when alone (0 peers) is a significant drain on the battery. Even when not alone, our devices keep uploading their location to the Web service because of variations in the RSSI from GSM cell towers and re-association with a different GSM cell tower, despite the nodes being static in this experiment. GSM localization that is more robust to such variations should help. In Figure 14, we show the advantage of using a notification system instead of polling. When there are no other devices around, the savings are tremendous – lifetime increases from 15.7 hours to 35 hours. Since there are no devices around, the device keeps Wi-Fi off and does not scan over Bluetooth, and does not need to poll the service over 3G.

We now evaluate the improvement offered by the back-light optimization from § 4.3. The previous experiments do not use this optimization because we lack accurate usage models of our application. Hence in Table 3, we present an evaluation of this optimization based on emulation of the Reality Mining data [44]. The data covers a large number of users across many days and indicates when their phones are idle versus in use. We pick 10 users at random and focus on their behavior for a random day. For periods of time when all the devices are idle, we follow our technique from § 4.3 and keep Wi-Fi off and do not scan on Bluetooth. We repeat these emulations multiple times by picking 3 different days at random, and 3 different sets of 10 users, and present average numbers in Table 3. While this emulation may not perfectly match real usage, these results show that this optimization has the potential to increase lifetime by 30%.

Finally, we present Table 4 where we evaluate the energy consumption of Virtual Compass as we vary the density of deployment. The lifetime does not significantly vary with density. There is a slight increase in lifetime as density decreases, and this is because there are fewer peers that are directly reachable over Bluetooth, and hence fewer connections need to be setup to measure RSSI.

Latency: Latency is another important metric – Virtual Compass should sense changes in the neighbor graph fast enough for applications that want to detect social interactions, and for those that provide maps in real-time to users. Figure 15 shows the overhead of different components of Virtual Compass. Bluetooth scanning is particularly slow, and we discuss this in more detail in § 7. Bluetooth pairing is needed to work around a limitation of the Bluetooth interface in Windows Mobile. The Windows Vista Bluetooth stack does pass up

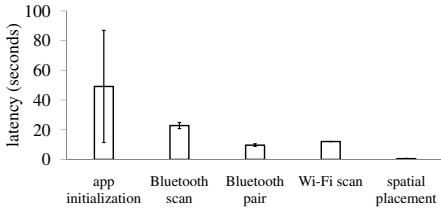


Fig. 15. This graph shows the latency of various tasks in Virtual Compass, along with error bars indicating variance across several runs. A total of 10 stationary devices were placed in a $100m^2$ area. “app initialization” is dominated by communication with the Facebook Web site.

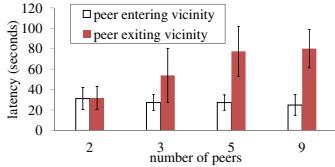


Fig. 16. This graph shows the latency of Virtual Compass detecting a peer moving into or moving out of the vicinity, with different numbers of nearby, stationary peer devices

Table 5. This figure shows the stability of the neighbor graph when using just Bluetooth, just Wi-Fi, or both. We placed 2 devices 10m apart, and ran experiments for 2 hours.

System	Neighbor graph stability
Bluetooth	14%
Wi-Fi	90%
Virtual Compass	94%

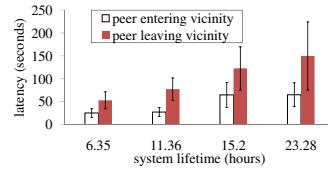


Fig. 17. This graph shows the trade-off between the latency of detecting a peer moving into or moving out of the vicinity, against the lifetime of Virtual Compass . We varied the Wi-Fi scan interval between 30 seconds and 4 minutes.

RSSI values from a Bluetooth scan without having to pair and connect, and so we are confident that this problem is not inherent to Bluetooth.

The time taken to detect the arrival of a new peer depends not only on the latency of Wi-Fi and Bluetooth scans, but also on how reliable scanning is. In Table 5, we present the probability of finding a peer device with a Bluetooth scan, Wi-Fi scan and both. Bluetooth is particularly poor because when two adjacent devices are scanning (and hence frequency-hopping) simultaneously, the probability of both being on the same channel and hence discovering each other is very low. This problem is specific to Bluetooth, as the stability of Wi-Fi is much higher. Since Virtual Compass uses both radios, it can detect the presence of a peer device more reliably than either alone.

We now evaluate how quickly Virtual Compass detects peer movement. In particular, we consider: (1) time elapsed between a peer entering the vicinity of a device and the peer showing up on the map, and (2) time elapsed between a peer leaving the vicinity and it disappearing from the map. We evaluate both

latencies in Figure 16. The latency of detecting a new peer is dominated by the frequency of scanning – in steady state, Bluetooth scanning occurs once every 10 minutes, but Wi-Fi occurs every minute. Since the graph shows the average across many runs, the average latency for detecting a new peer is 30 seconds, because of Wi-Fi scanning. Peer departure can be a higher latency operation as the number of peers increases because all peers have to remove the exiting peer from their neighbor graph, else it will still appear in the map due to multi-hop discovery. Hence Bluetooth’s slower scan time dominates peer departure latency.

Reducing the scan interval of Wi-Fi and Bluetooth can reduce latency, but it comes at the cost of energy. Figure 17 explores this trade-off. The second set of bars at 11.36 hours corresponds to the 5 peers bars from Figure 16. We can double the lifetime to 23.28 hours at the cost of doubling latency. However, halving the lifetime to 6.35 hours does not significantly reduce latency. Hence we believe that our choice of the Wi-Fi wake-up and scan interval of 1 minute and the Bluetooth limits of 10 seconds to 10 minutes offer the best trade-off.

7 Discussion

We now discuss performance optimizations for Virtual Compass.

Improving accuracy: While Virtual Compass uses a single RSSI-distance profile, we could use different profiles for different environments, such as outdoors versus indoors. This would require a mobile device to determine if it is outdoors, and then apply the corresponding RSSI-distance profile. We are exploring two ways to solve the problem of detecting that the user is outdoors. First, if a GPS signal is available, then we can assume the user is outdoors. Second, we can use user feedback.

Reducing Latency: Virtual Compass’s latency in detecting node movement is significantly impacted by Bluetooth scanning. Two devices that simultaneously scan over Bluetooth can miss each other because each may use a different frequency hopping sequence such that the two devices never end up on the same channel at the same time. To alleviate this problem, we are investigating certain Bluetooth 1.2 chipsets that allows *enhanced inquiry* which is supposed to make discovery reliable and fast (less than 5 seconds).

Reducing energy consumption: Not all mobile devices have similar energy budgets. A laptop has a larger battery than a mobile phone. Furthermore, some environments may have desktops with wireless interfaces. We posit that it is beneficial for mobile phones to offload the task of aggressively scanning for device movement to nomadic infrastructure that is energy rich. The nomadic infrastructure can scan very frequently, and if it detects that a new device has come into range, or a device has moved or left, then it can signal other devices to scan and re-compute the neighbor graph. We are presently investigating schemes for efficiently offloading computation to more powerful infrastructure.

Privacy and security: There are privacy and security issues that we have not addressed in Virtual Compass. In our current implementation, a user’s numeric Facebook ID is her mobile device’s ID in peer localization. In our application,

she only sees her Facebook friends. However, our underlying peer localization component has a complete map of all the devices in the vicinity. A wily user could potentially misuse this information. As a solution, we could use a periodically changing random number for the device ID. Each device would register this ID with an applet on Facebook. Any device that wants to discover the user identity will have to query the applet, which can verify if that user is a friend.

8 Conclusion

Most of today's mobile social applications use *absolute* location to locate nearby peers, which is often difficult to obtain with reasonable accuracy in indoor environments. In this paper, we describe Virtual Compass, a peer-based localization system for mobile phones, which provides *relative* positioning by placing peers in a 2D plane without requiring any infrastructure support. Virtual Compass enables many emerging mobile applications that want the ability to sense social interactions: it provides the distance between different people which can then be combined with external information about those people's social relationships. A key area of future work is to use this information to build applications that automatically infer of social context of such interactions.

Virtual Compass leverages the multiple radios available on today's smartphones to provide the accuracy needed for the above applications. It uses several energy management techniques that frugally use radios without compromising location accuracy. We have implemented Virtual Compass for Windows Mobile phones. We have implemented a simple application, FriendMeter, which uses Virtual Compass to sense the distances between a user and her Facebook friends who are in the vicinity. We evaluate Virtual Compass on a nine node testbed, and our results show that it places a device with an average distance error of only 1.9 meters. Virtual Compass's energy management algorithms produce a battery lifetime that is four to nine times that of a device that does not use sophisticated energy management to provide peer localization.

Acknowledgements. We would like to thank our shepherd, Alexander Varshavsky, and the anonymous reviewers for their useful feedback. Part of this work was supported under awards NSF CNS-0519881, NSF CNS-0447877, and DARPA HR0011-09-1-0020.

References

- [1] Miluzzo, E., Lane, N.D., Fodor, K., Peterson, R.A., Lu, H., Musolesi, M., Eisenman, S.B., Zheng, X., Campbell, A.T.: Sensing meets mobile social networks: The design, implementation and evaluation of the CenceMe application. In: SenSys (2008)
- [2] Loopt, <http://loopt.com>
- [3] Dodgeball Social Networking, <http://dodgeball.com/>
- [4] Varshavsky, A., de Lara, E., Hightower, J., LaMarca, A., Otsason, V.: GSM indoor localization. Pervasive and Mobile Computing Journal (December 2007)

- [5] LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, J., Sohn, T., Howard, J., Hughes, J., Potter, F., Tabert, J., Powledge, P., Borriello, G., Schilit, B.: Place Lab: Device Positioning using radio beacons in the Wild. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) PERVASIVE 2005. LNCS, vol. 3468, pp. 116–133. Springer, Heidelberg (2005)
- [6] Haeberlen, A., Flannery, E., Ladd, A., Rudys, A., Wallach, D., Kavraki, L.: Practical robust localization over large-scale 802.11 wireless networks. In: MobiCom (2004)
- [7] Maroti, M., Kusy, B., Balogh, G., Volgyesi, P., Nadas, A., Molnar, K., Dora, S., Ledeczi, A.: Radio interferometric geolocation. In: SenSys (2005)
- [8] LaMarca, A., de Lara, E.: Location systems: An introduction to the technology behind location awareness. Synthesis Lectures on Mobile and Pervasive Computing (2008)
- [9] Laasonen, K., Raento, M., Toivonen, H.: Adaptive on-device Location Recognition. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 287–304. Springer, Heidelberg (2004)
- [10] Laitinen, H., Lahteenmaki, J., Nordstrom, T.: Database Correlation method for GSM Location. In: VTC (2001)
- [11] Ekahau Wi-Fi-based Real-time Tracking and Site Survey Solutions, <http://ekahau.com>
- [12] Bahl, P., Padmanabhan, V.N.: RADAR: An in-building RF-based User Location and Tracking System User Location and Tracking System. In: INFOCOM (2000)
- [13] Borriello, G., Liu, A., Offer, T., Palistrant, C., Sharp, R.: WALRUS: Wireless Acoustic Location with Room-Level Resolution Using Ultrasound. In: MobiSys (2005)
- [14] Priyantha, N.B., Chakraborty, A., Balakrishnan, H.: The Cricket Location-Support System. In: MOBICOM (2000)
- [15] Hopper, A., Harter, A., Blackie, T.: The Active Badge System. In: InterCHI (1993)
- [16] Versus Technologies, <http://versustech.com>
- [17] Ubisense, <http://ubisense.net>
- [18] Pantopic Social Networking, <http://pantopic.com/>
- [19] Rummble Social Networking, <http://rummble.com/>
- [20] Holmquist, L.E., Falk, J., Wigström, J.: Supporting group collaboration with interpersonal awareness devices. Journal of Personal Technologies 3, 13–21 (1999)
- [21] Krumm, J., Hinckley, K.: The NearMe Wireless Proximity Server. In: Davies, N., Mynatt, E.D., Siio, I. (eds.) UbiComp 2004. LNCS, vol. 3205, pp. 283–300. Springer, Heidelberg (2004)
- [22] Peng, C., Shen, G., Zhang, Y., Li, Y., Tan, K.: Beep Beep: A High Accuracy Acoustic Ranging System Using COTS Mobile Devices. In: SenSys (2007)
- [23] Bluehoo, <http://bluehoo.com>
- [24] Li, K.A., Sohn, T.Y., Huang, S., Griswold, W.G.: Peopletones: A system for the detection and notification of buddy proximity on mobile phones. In: MobiSys (2008)
- [25] Hazas, M., Kray, C., Gellersen, H., Agbota, H., Kortuem, G., Krohn, A.: A Relative Positioning System for Co-located Mobile Devices. In: MobiSys (2005)
- [26] Holmquist, L., Falk, J., Wigstrom, J.: DOLPHIN: A Practical Approach for Implementing a fully Distributed indoor Ultrasonic Positioning System. In: Davies, N., Mynatt, E.D., Siio, I. (eds.) UbiComp 2004. LNCS, vol. 3205, pp. 347–365. Springer, Heidelberg (2004)

- [27] Zhong, Z., He, T.: MSP: Multi-Sequence Positioning of Wireless Sensor Nodes. In: SenSys (2007)
- [28] Bulusu, N., Heidemann, J., Estrin, D.: GPS-Less Low-Cost Outdoor Localization for Very Small Devices. IEEE Personal Communications (2000)
- [29] Guha, S., Murty, R., Sirer, E.G.: Sextant: a unified node and event localization framework using non-convex constraints. In: MobiHoc (2005)
- [30] Varshavsky, A., Pankratov, D., Krumm, J., Lara, E.D.: Calibree: Calibration-free Localization using Relative Distance Estimation. In: Indulska, J., Patterson, D.J., Rodden, T., Ott, M. (eds.) PERVASIVE 2008. LNCS, vol. 5013, pp. 146–161. Springer, Heidelberg (2008)
- [31] Zhang, P., Martonosi, M.: Locale: Collaborative localization estimation for sparse mobile sensor networks. In: IPSN (2008)
- [32] Moore, D., Leonard, J., Rus, D., Teller, S.: Robust distributed network localization with noisy range measurements. In: SenSys (2004)
- [33] Stoleru, R., He, T., Stankovic, J.A., Luebke, D.: High-accuracy, low-cost localization system for wireless sensor network. In: SenSys (2005)
- [34] Dabek, F., Cox, R., Kaashoek, F., Morris, R.: Vivaldi: A decentralized network coordinate system. In: SigComm (2004)
- [35] Zanca, G., Zorzi, F., Zanella, A., Zorzi, M.: Experimental comparison of rssi-based localization algorithms for indoor wireless sensor networks. In: REALWSN (2008)
- [36] Chandra, R., Padhye, J., Wolman, A., Zill, B.: A Location-based Management System for Enterprise Wireless LANs. In: NSDI (2007)
- [37] Banerjee, N., Agarwal, S., Bahl, P., Chandra, R., Alec Wolman, M.C.: Virtual compass: relative positioning to sense mobile social interactions. Technical report (2009)
- [38] McAllester, D.: The Rise of Nonlinear Mathematical Programming. In: ACM Computing Surveys (1996)
- [39] Gaonkar, S., Li, J., Choudhary, R.R., Cox, L., Schmidt, A.: Micro-Blog: Sharing and Querying Content Through Mobile Phones and Social Participation. In: MobiSys (2008)
- [40] Nokia Nokoscope Data obtained via private communication
- [41] Agarwal, Y., Chandra, R., Wolman, A., Bahl, P., Chin, K., Gupta, R.: Wireless wakeups revisited: energy management for VoIP over Wi-Fi smartphones. In: MobiSys (2007)
- [42] Chandra, R., Padhye, J., Ravindranath, L., Wolman, A.: Beacon-Stuffing: Wi-Fi without Associations. In: HotMobile (2007)
- [43] Microsoft Azure SQL Data Service, <http://microsoft.com/azure/data.mspx>
- [44] Reality Mining Dataset, <http://reality.media.mit.edu/>

The Geography of Taste: Analyzing Cell-Phone Mobility and Social Events

Francesco Calabrese¹, Francisco C. Pereira^{1,2}, Giusy Di Lorenzo¹, Liang Liu¹, and Carlo Ratti¹

¹ MIT Senseable City Laboratory, Cambridge, MA

² Centro de Informatica e Sistemas da Universidade de Coimbra, Coimbra, Portugal
`fcalabre@mit.edu, camara@dei.uc.pt, giusy@mit.edu, liuliang@mit.edu`

Abstract. This paper deals with the analysis of crowd mobility during special events. We analyze nearly 1 million cell-phone traces and associate their destinations with social events. We show that the origins of people attending an event are strongly correlated to the type of event, with implications in city management, since the knowledge of additive flows can be a critical information on which to take decisions about events management and congestion mitigation.

1 Introduction

Being able to understand and predict crowded events is a challenge that any urban manager faces regularly, particularly in big cities. When it is not possible to determine the exact numbers (e.g., from ticket sales), the typical approach is based on intuition and experience. Even when the exact number of event attendees is known, it is still difficult to predict their effect on the city systems when traveling to and from the event. During the last years, the Pervasive Computing community has developed technologies that now allow us to face the challenge in new ways. Due to their ubiquity, GSM, bluetooth or WiFi localization technologies such as in [1][2][3] can now be explored at a large scale.

The development of methodologies that allow for an accurate characterization of events from anonymized and aggregated location information has further potential implications for Pervasive Computing research, namely enhancing the context awareness. Location based services can be imagined that take into account the predicted effect of events in the city. For example, navigation systems that try to avoid the predicted congested areas, social applications that lead people to (or away from) the “crowds” or interactive displays that adapt to the expected presence of people. Other applications could include inference of points of interest or emergency response planning.

In this paper, we present our work on the combination of analysis of anonymized traces from the Boston metropolitan area with a number of selected events that happened in the city attracting considerably sized crowds. The objective is to characterize the relationship between events and its attendees, more specifically of their home area. The hypothesis is that different kinds of events bring people

from different areas of the city according to distribution patterns that maintain some degree of constancy. The rationale is that people maintain regular patterns of preferences throughout time (e.g., a sports fan will often go watch games; a family that has children will often go to family events). While we make no assumptions on the distributions of “types of people” among areas of a city, it is reasonable to assume that aggregate patterns of “types of neighborhoods” will emerge.

The next section is dedicated to further understanding the motivation and context of this work, followed by a review of related work. The explanation of the data involved in this study is then made in section 4 while the core of the paper is presented in section 5, where we present our methodology and experimental results.

2 Motivation

In 2008, a study from the U.S. Federal Highway Administration [4] was dedicated to investigate the economic and congestion effects of large planned special events (PSEs) on a national level. The clearer understanding of the scale of PSEs and their economic influence is essential to achieve a more efficient transportation planning and management of traffic logistics of such events. In that study, the authors find that there are approximately 24,000 PSEs annually with over 10,000 in attendance across USA, or approximately 470 per week. These numbers, possibly similar in other parts of the world, call for application of efficient techniques of crowd analysis. From the point of view of Pervasive Computing, besides the very task of analyzing *digital footprints* obtained from ubiquitous devices, which lies in the crux of this research, other questions arise that transcend this area.

One question is understanding the stability of crowd patterns in medium to large scale events. If regularity is confidently demonstrated, then pattern sensitive services can be developed that improve the events experience (e.g. providing mobility advisory for evacuation after the event). The converse question is also relevant, namely the characterization of different neighborhoods by knowing what kinds of events their residents prefer to attend. This would allow for the construction of emotional/hobby maps of each block, becoming in turn contextual information about space, adding value to location aware systems.

Perhaps the most obvious problems at the local scale and those that we will illustrate in this paper comprise one-off spatial events which involve the movement of large numbers of people over short periods of time. These largely fall within the sphere of entertainment although some of them relate to work, but all of them involve issues of mobility and interaction between objects or agents which generate non trivial problems of planning, management, and control. The classic example is the football match but rock concerts, street parades, sudden entry or exit of crowds from airports, stations, subway trains, and high buildings could be included. Particularly these types of event, however, have tended to resist scientific inquiry, and have never been thought to be significant in terms of their impact on spatial structure, or to be worthy of theory.

3 State of the Art

Before describing the related work, we bring some definitions that collect relative agreement in the literature. Within the topic of *crowd analysis*, we consider *event inference* and *crowd modeling*. The detection of an existence of a crowd given available data (e.g. images about a place, aggregated communications) is the objective of event inference. Such event may or may not be predictable or correspond to an actual public *special event*. The task of crowd modeling consists of building patterns or descriptions of (a) crowd(s) that enable prediction or simulation of crowd behaviour. A successful crowd model allows for useful applications such as predicting the use of a space, planning accessibility, preventing dangerous situations or planning an emergency evacuation, for example.

Following [5,6] we propose to organize crowd modeling according to three levels: microscopic, macroscopic, mesoscopic. At the microscopic level, the individual is the object of study, while at the macroscopic level, we work with groups. The mesoscopic model combines the properties of the previous two, either keeping a crowd as a homogeneous mass but considering an internal force or keeping the characters of the individuals while maintaining a general view of the entire crowd [6].

From the point of view of data collection, the traditional approach consists of aggregating data from control points (e.g. number of tickets sold; nights in hotels, number of people per room; counting people) as well as from surveys provided to randomly chosen individuals (e.g. [7]). During the nineties, research from computer vision brought alternative (and non-intrusive) methods that allowed to extract crowd related features, namely on detecting density (quantity of people over space), location, speed and shape (e.g. [8]). Although such properties allow for useful analysis, they are restricted to the space of study (or spaces of study, depending on the number of cameras available).

The often mentioned outburst of mobile phones during late 20th century accompanied by the more recent trend of sensors and advanced communication systems (e.g. GPS, digital cameras, Bluetooth, WiFi) allow for unforeseen amounts of data from urban areas through which to study both groups [9,10,11], individuals [12] or both [3].

The afore mentioned technologies present different challenges and potential regarding event inference. The traditional methods are slow and precise when the event is controlled in space but with little precision in the opposite case (e.g. [7]). Computer vision allows for automatic inference of events also providing some properties such as those referred above but limited to areas with visual data (e.g. [8]). Using digital footprints such as communication or GPS traces, we can reach wider areas but with lower precision in comparison to these methods. In [13], the authors analyse the presence of tourists in a wide area (Lower Manhattan) during a public art installation (the “NYC waterfalls”) for 4 months using cell-phone activity. In the Reality Mining project, 100 students from the MIT campus carried smart-phones over 9 months and their social and individual behaviours were analysed using Cell ID and Bluetooth [3]. In a case study of tourism loyalty in Estonia, Ahas et al [14] show that the sampling and analysis

of passive mobile positioning data is a promising resource for tourism research and management. They show that this type of aggregated data is highly correlated with accommodation statistics in urban touristic areas. In a case study in Tawaf during the Hajj, Koshak and Fouda [1] verified how GPS and GIS data can be utilized to perform tempo-spatial analysis of human walking behavior in an architectural or urban open space.

In terms of level of detail, traditional methods are generally adequate for macroscopic detail (unless individualized data is collected), computer vision allows for any of the levels but is particularly suited to macro- and mesoscopic analysis while digital footprints can be useful for any of the levels discussed, namely microscopic when individual privacy is properly protected. Of course, the precision is dependent on the penetration rate of the technology of study (e.g. number of cell-phone users in the crowd).

As for modeling of crowd behaviour, related work can be found at several distinct fields. In computer vision, crowd models are built as representations of recurrent behaviours by analysing video data of the crowd through vision methods. In physics, many approaches have been built inspired by using fluid dynamics [2], swarms [6][7] or cellular automata [8]. In literature, there is no characterization of particular “special events” bounded in time and space and in general their goals are at the mesoscopic level (model group from aggregated individual modeling). Also, these studies of digital footprints have used aggregated information of people, rarely reaching the (anonymized) individual detail.

4 Data Description

The data analyzed corresponds to an area of 15×15 kilometers within Boston, as shown in Figure 1. This area includes the main event venues in the state of Massachusetts and some of the most densely populated residential areas of Greater Boston. We analyzed cellphone mobility and events happening in that area for the period from July 30th to September 12th of 2009, as we describe next.

4.1 Cellphone Mobility Data

The dataset used in this project consists of anonymous cellular phone signaling data collected by AirSage[2], which turns this signaling data into anonymous locations over time for cellular devices. This aggregated and anonymous cellular device information is used to correlate, model, evaluate and analyze the location, movement and flow of people in the city. The dataset consists of 130 millions of anonymous location estimations - latitude and longitude - from close to 1 million devices (corresponding to a share of approximately 20% of the population, equally spread over space) which are generated each time the device connects to the cellular network, including:

- when a call is placed or received (both at the beginning and end of a call);
- when a short message is sent or received;

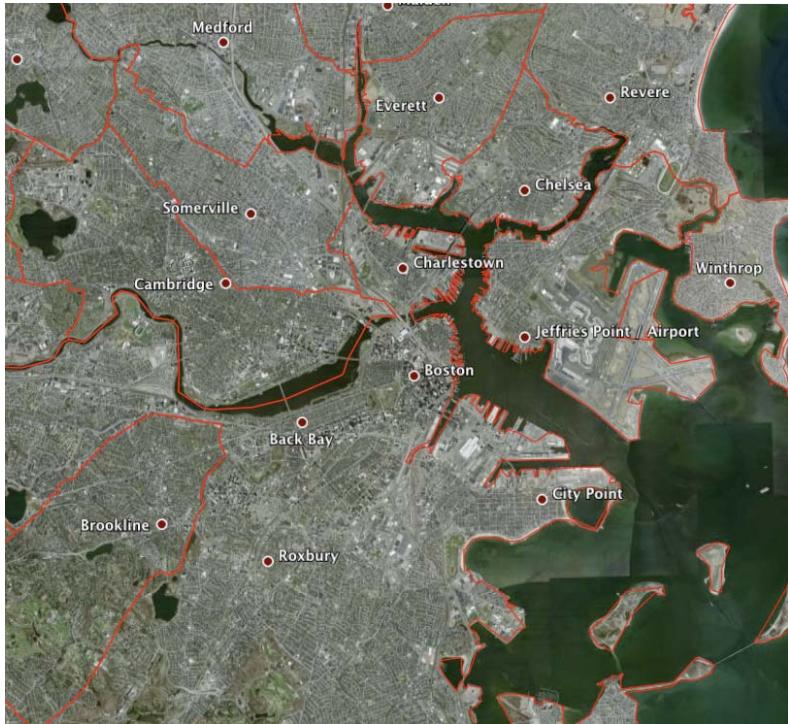


Fig. 1. Study area

- when the user connects to the internet (e.g. to browse the web, or through email programs that periodically check the mail server).

Since the location measurements are generated based on signaling events, i.e. when the cellphone communicates with the cell network, the resulting traces are far from regularly sampled. Besides, cellphone-derived location data has a greater uncertainty range than GPS data, with an average of 320 meters and median of 220 meters as reported by AirSage [2] based on internal and independent tests.

4.2 Events Data

Events in the Boston metropolitan area were selected to evaluate whether people from different areas of the city chose to attend different types of events. For the selection of events, it was important to find the largest set that occurs during the time window of the study and that complies with a number of requirements:

- The attendance should have relevant size in order to allow for a significant number of identified users.
- Be isolated in space with respect to neighboring events. To avoid ambiguity in the interpretation of results, we decided to give a minimum margin of one kilometer in any direction to any other large size simultaneous event.

- The venue of the event should correspond to a well defined area with considerable dimensions. It is also important to minimize the potential to misinterpret people staying in other places for event attendees (e.g. staying in a restaurant nearby).
- Be isolated in time to any other big event (i.e. not be in the same day). For a proper analysis, it is also important to guarantee that the statistics of presence (or absence) of people in the events is minimally dependent on external events as this would lead to erroneous conclusions.
- Have a duration of at least 2 hours. The assumption is that attendees are at the venue specifically for the event. With small time durations, it becomes difficult to distinguish occasional stops from actual attendance.

Our goal was to reduce the influence of dependencies between different events and the ambiguity in determining whether a person is attending an event or simply staying in a place near. Another concern was to select events from a variety of categories, namely Performance Arts, Sports events, Family events, Music and Outdoor Cinema.

We analyzed the Boston Globe event website [I9] and selected 6 different venues, corresponding to a total of 52 events. We also contacted the organizers of some events in order to get their attendance estimations. In Table II, we show a summary of the events.

It is notable that two of the cases violate one or more of the requirements, namely indoor cinema in the Museum of Science at the same time as the cinema sessions in the Hatch Shell and with an intersection with the Children's museum event. The Cirque du Soleil event also conflicts with the summer concerts at the Hatch Shell. The reason is that, since the venues are far apart and only one has space for very large crowds (Hatch Shell), the overall results should not be affected. In figure 2, we show the event locations.

Table 1. Event list

Venue	Events	Type	Date	Time
Fenway Park	11 Red Sox games (baseball)	Sports	10, 11, 12, 25 and 26 Aug, 7-10pm 8, 9 September	
Agganis Arena	Cirque du Soleil Alegria (2 times)	Performance Arts	26, 27 of Aug.	7:30-10pm
DCR Hatch Shell	Friday flicks (5)	Cinema	31 July, 7, 14, 21 and 28 August	8-10pm
DCR Hatch Shell	Summer concerts (5)	Music	5, 12, 29 and 26 August, 2 September	7-9pm
Museum of Science	Friday nights (7)	Cinema	31 July, 7, 14, 21 and 28 August, 4 and 11 Sep.	5-9pm
Boston Common	Shakespeare on the Boston Common (15)	Performance Arts	31 July, 1, 2, 4-9, 11-16 August	8-10pm
Children's museum	Target fridays (7)	Family	31 July, 7, 14, 21 and 28 August, 4 and 11 Sep.	5-9pm



Fig. 2. Event locations

4.3 Data Preparation

The data as provided does not directly allow determining mobility traces of users. We then applied a process to perform an estimation of the mobility choices each user takes over time. The process involves two steps:

- Inferring what we call *stops*: places in which a person has stopped for a sufficiently long time.
- Inferring the home location of each user.
- Performing a spatio-temporal analysis of the sequence of stops to detect which users are attending a given event.

In order to infer the sequence of stops that each user makes, we first characterized the individual calling activity and verified whether that was frequent enough to allow monitoring the user's movement over time with fine enough temporal resolution. As we said in the section 4.1, each location measurement m_i , collected for every cellphone, is characterized by a position p_i , expressed in latitude and longitude, and a timestamp, t_i . For each user we measured the interevent time i.e. the time interval between two consecutive network connections (similar to what was measured in [20]). The average interevent time measured for the whole population is 260 minutes, much lower than the one found in [20]. Since the distribution of interevent times for an individual spans over several decades, we further characterized each calling activity distribution by its first and third quantile and the median. Fig. 3 shows the distribution of the first and third quantile and the median for the whole population. The arithmetic average of the medians is 84 minutes (the geometric average of the medians is 10.3 minutes) which results small enough to be able to detect changes of location where the user stops as low as 1.5 hours (time comparable to the average length of the considered social events).

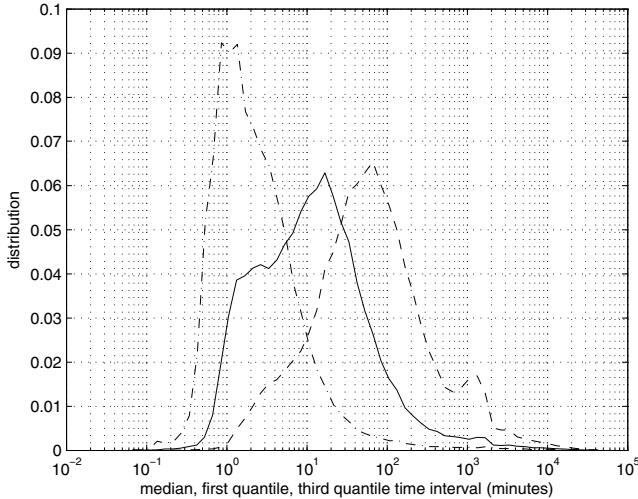


Fig. 3. Characterization of individual calling activity for the whole population. Median (solid line), first quantile (dash-dotted line) and third quantile (dashed line) of individual interevent time.

The analysis above tells us that the cellphone data can be used to extract users' movements as it changes over the course of the day. To extract the sequence of stops, we first extracted trajectories from the individual location measurements. A trajectory is a sequence of chronological locations visited by a user.

$$Traj = \{p_1 \rightarrow p_2 \rightarrow \dots \rightarrow p_n\}$$

A sub-trajectory is obtained by segmenting the trajectory with a spatial threshold ΔS , where $distance(p_i, p_{i+1}) > \Delta S, i = 1..n$. The segmentation aims at removing spatial gaps between two recorded points (p_i, p_{i+1}) of more than ΔS . If a gap is found, p_i becomes the end point of the last sub-trajectory, and p_{i+1} becomes the starting point of the new sub-trajectory. Once sub-trajectories are detected, we first resampled with a constant sampling time T_c and then applied to them a low pass filter in order to eliminate some measurement noise contained in the data (as done in [21] [22]). For each sub-trajectory we determined the time at which the user stops traveling, and call the location stop s .

The extraction of a stop depends on two parameters: time distance threshold (T_{th}) and a spatial distance threshold (S_{th}). Therefore, a single stop s can be regarded as a virtual location characterized by a group of consecutive location points

$$P = \{p_s, p_{s+1}, \dots, p_m\},$$

where $\forall s \leq i, j \leq m, max(distance(p_i, p_j)) < S_{th}$ and $t_m - t_s > T_{th}$.

Once the stops have been extracted, the home location of each user is then estimated as the most frequent stop during the night hours.

The information about the stops and home location allows us to derive the mobility choices of users, and detect whether they are attending an event, and the origin of the trip to attend the event.

Hence, we first grouped together users that live close in space (their home location is close), creating a grid in space where the side of each cell is 500 meters. Then, to understand if a user is attending an event we checked the following assumptions: i) the user stops in the same cell of the event location, ii) the stop overlaps at least 70 percent with the duration of the event, and iii) the user's home location is different from the event location. The Figure 4 shows the idea behind these assumptions. We do not require a full overlap to take into account the fact that we are not able to detect locations of users with a very high frequency, and so might not consider users just because they do not connect to the network at the beginning and end of the event.

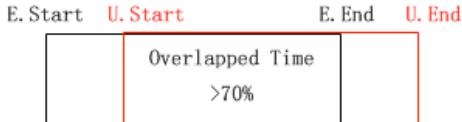


Fig. 4. Audience detection algorithm: if intersection of duration of user stop and duration of the event is greater than 70 percent and user's home is not the same as the event location, then we mark the user as audience of the event

Finally, the mobility choices are derived by inferring the spatial origins' distribution of the people that attempt to the events. Given an event, for each cell of the grid we count the number of people attending to that event and whose home location falls inside that cell. This spatial distribution can then be plot on a map to show the areas of the city which are more interested in attending the event. Examples of such map are shown in the following section.

5 Methodology

Our methodology for describing events through mobility choices is based on the use of the estimated origins of people attending to the events. Figure 5 shows some examples of spatial variation of the estimated origins of people attending different events.

Sport events such as baseball games (Figure 5(a)) attract about double the number of people which normally live in the Fenway Park area. Moreover, those people seem to be predominantly attended by people living in the surrounding of the baseball stadium, as well as the south Boston area (Figure 5(b)).

Performing arts events such as the "Shakespeare on the Boston Common" (Figure 5(c) and 5(d)) which his held yearly, attract people from the whole Boston metropolitan area, and very strongly people which live in the immediate surroundings of the Boston Common (average distance lower than 500 meters).

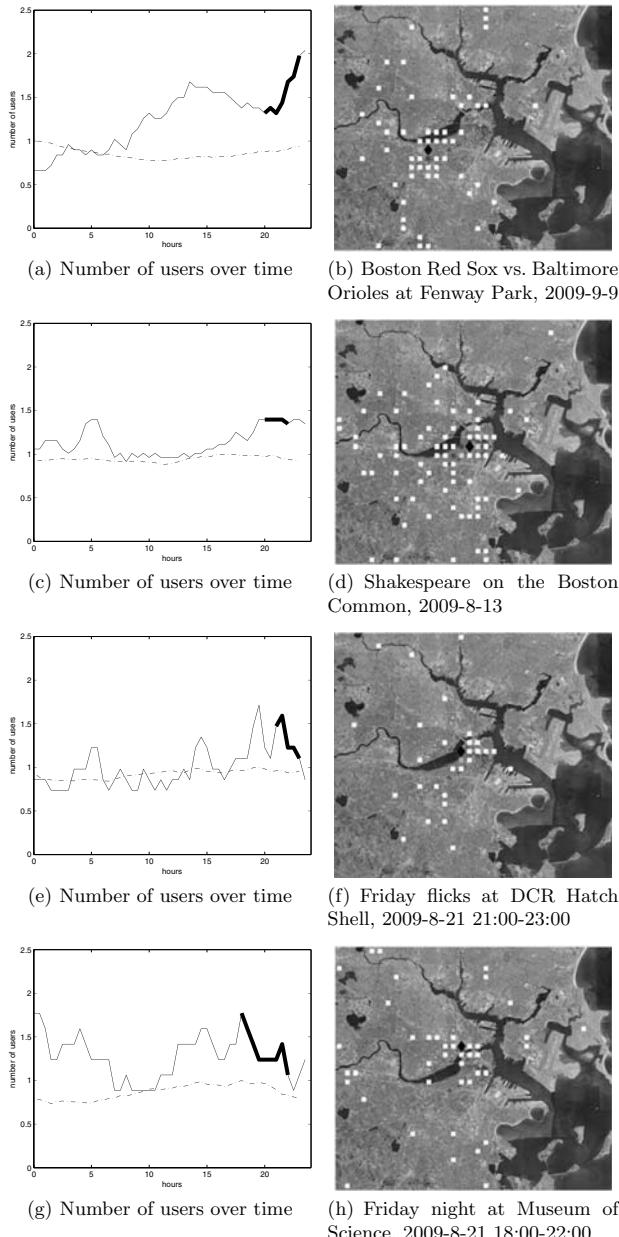


Fig. 5. Examples of events in Boston. Figures *a*, *c*, *e* and *g* show the number of users at the locations of events over the course of the day of the event (solid line) compared to an average day (dash-dotted line). Note that number of users are scaled with respect to the maximum in an average day. Figures *b*, *d*, *f* and *h* show the locations of the events (diamond) and estimated origins distribution of people attending the events: shade from light (low) to dark (high).

The number of people attending the event is instead about 1.5 times greater than what it is usually found in the Boston Common.

By comparing the two images in Figures 5(b) and 5(d) it is easy to understand that most of the people attending to one type of event are most probably not attending the other type of events, showing a complementary role of sports and arts events in attracting different categories of people.

Finally, Figures 5(e), 5(f), 5(g), 5(h) show the spatial distribution of origins of people for two events (movie screening) happening almost at the same time in two very close areas in Boston (DCR Hatch Shell and Museum of Science).

Since the origins of people attending an event are strictly related to the location and type of events, we argue that by using just this information we would be able to predict the type of event. If a relationship between origin of people and type of event is found, it would be possible to determine the abnormal and additive travel demand due to a planned event by just considering the type of that event. It would then be possible to provide a city with critical information on which to take decisions about changes in the transportation management, e.g. increasing the number of bus lines connecting certain areas of the city to the venue of the event.

In the next section we will show 8 different models that we have developed to perform the prediction of the type of event starting from the mobility data associated with it.

Note that the number of attendees we are able to detect is strictly related to the share of the telecom operator partnering with Airsage. We empirically tested that the number of users correspond to about 20% of the population (as reported by the latest US census) and is equally distributed over the different zipcodes. Since we selected only events with relevant size, this allowed us to detect significant numbers of users per event. We verified that those numbers are also consistent for events of the same type, proving that there is a significant and consistent number of detected attendees allowing us to perform the comparative analysis reported in the next section. Estimating the actual number of attendees is still an open problem, considering also that ground truth data to validate models is sometime absent or very noisy (usually based on head counts or aerial photography).

5.1 Prediction

The task at hand is to understand the relationships between events and origins of people. Particularly, we seek for the predictive potential of events in respect to mobility phenomena. This can be seen from two perspectives: a classification task in which we want to understand how a vector of features (e.g., attendees origin distribution) predicts a classification (e.g., an event name or type); a clustering task, in which the feature vectors are distributed according to similarity among themselves.

We used the Weka open source platform [23], which contains a wide range of choices for data analysis. For classification, we use a Multilayer Perceptron, with one hidden layer and the typical heuristic of (*classes+attributes*)/2 for the

number of nodes. For clustering, we apply the K-Means algorithm (with $K = \#$ event types or $K = \#$ event places). In each experiment, we used 10-fold cross-validation, in which a tenth of the dataset is left aside for testing the algorithm while using the remaining for training. This train-test process is ran 10 times (one for each tenth of the dataset).

6 Experiments

We aggregated attendees in terms of zipcode area and distance to event, discretized in 2000 bins. We did so because if we were to use a geographic coordinate of individuals, the resulting data would be sparse. Instead, by aggregating data geographically, we could find useful patterns. To avoid the strong bias towards attendees in the neighborhood of the event, we also remove those that live in the same area of the event (their home location falls in the same 500m x 500m cell of the event) because we would not be able to distinguish between event and home.

For each event, we created an *instance* that contains the corresponding attendee *origin pattern distribution*, evaluated at the level of the zipcode area (with average size of $4.5km^2$). For example, for one showing of the Shakespeare’s “Comedy of Errors” at the Boston Common, we have 96 attendees (users monitored by the system, with a share of about 20% of the population) and then count the total number of people coming from each zipcode.

Our goal is to test whether similar events show similar geographical patterns. More specifically, given *origin pattern distribution*, the goal is to predict the type of event (as defined in Table I).

We met this goal by testing 8 prediction models, and we measure their accuracy in terms of fraction of correctly identified event types.

Before training our algorithms, we analyzed the overall distribution of events to get the *classifier* baselines. The principle is to know the accuracy of a classifier that simply selects randomly any of the 5 event types or that always chooses the same event type, and use them as a baseline to compare for the improvement of the quality. The average value of this baseline is 23.34% (standard deviation of 4.03) for random classification. Differently, if the classifier chooses the event with highest probability (performing arts), the accuracy will be 35%.

The first experiment was to use all vectors as just described, applied to a Multilayer Perceptron. The result is a surprising 89.36% of correctly classified events in the test set. From the clustering analysis, we see that mostly attendees come from the event’s zipcode area, suggesting that people who live close to an event are preferentially attracted by it. To focus on effects other than close proximity, we created a new prediction model considering only people coming from zipcode different from the event’s.

The result is 59.57%, which still indicates the recurrence of origin patterns for events of the same type. A clustering analysis brings the distributions that we can see in Figure 6.

Further analyses were made by putting a minimum threshold of at least 10 attendees for each zipcode area and by using home-event distance instead of

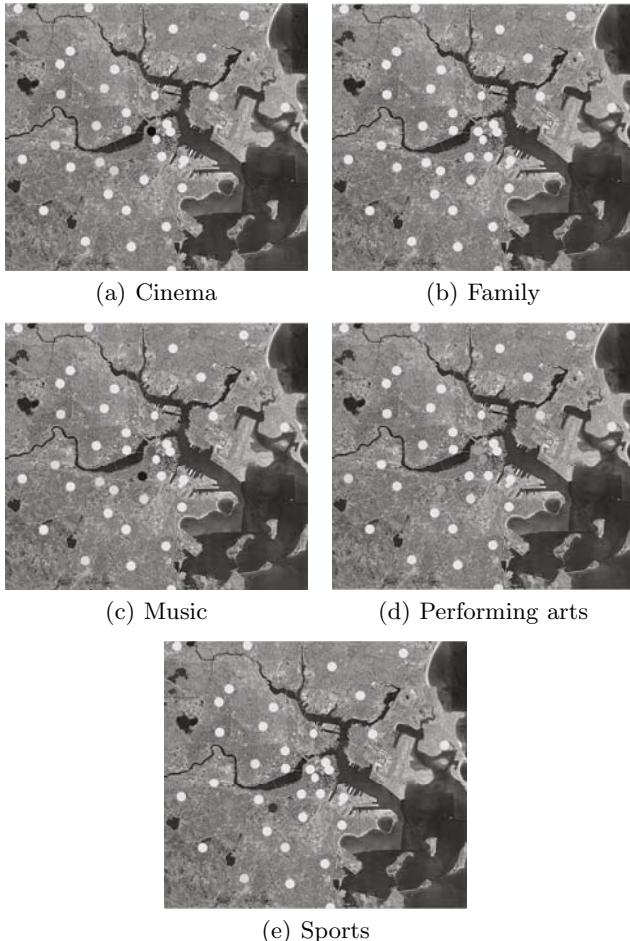


Fig. 6. Spatial visualization of clusters centroids. The circles correspond to the zipcode areas with value greater than zero. The shade from light (low) to dark (high) is proportional to the value.

zipcode (distance discretized in 2000 meter bins). The overall process of feature selection and attendee aggregation is the same as described above, and Table 2 shows the results. The item “Improvement” corresponds to the difference to the best baseline (fixed).

A first aspect that easily comes out of the predictions performed, is the clear difference between our classifiers and the baselines, indicating a consistency in the patterns found.

By comparing the results of the two predictions made using the zipcode areas, it is clear that the improvements found are consistent, and do not depend on small number of attendees that can be found sometimes in some zipcode areas.

Table 2. Summary of prediction results

Features	All attendees		Exc. event zipcode		Observation
	Precision	Improv.	Precision	Improv.	
Fixed baseline	35%				Always choose same class
Random baseline	23.34%				Random choice
Zipcode	89.36%	54.36%	59.57%	24.57%	All attendees
	95.74%	60.74%	53.19%	18.19%	All attendees when count>10
Distance	51.06%	16.06%	48.9%	13.9%	All att. Resolution 2000m

Interesting conclusions can be taken by comparing the improvement of the models using zipcode and distance. In fact the lower improvement shows that not only distance affects the event choices of people, but also where they live.

6.1 Limitations

Our methodology has two limitations. The location data is not continuously provided but is available only when users are active (call, SMS, data connection). This results in narrowing down the number of users we can analyze.

Secondly, we assign origins to users' home locations regardless of where their trips start. This does not hinder our analysis because we are interested in characterizing the taste of the local communities.

Further studies considering larger datasets of events and cell-phone users should be performed to obtain more statistically significant results.

7 Conclusions

Based on our analysis of nearly 1 million cell-phone traces we correlated social events people go to with their home locations. Our results show that there is a strong correlation in that: people who live close to an event are preferentially attracted by it; events of the same type show similar spatial distribution of origins. As a consequence, we could partly predict where people will come from for future events.

In the future, we will run the same study on datasets of cities other than Boston to verify to which extent the city's individual characteristics affect the patterns found.

Explicit spatial knowledge about crowd environment could also be considered to improve the proposed model.

Acknowledgements

Acknowledgments go to Airsage for proving us with the data, and to Leonardo Soto, Daniele Quercia, Mauro Martino, Carlo Ratti, Assaf Biderman for their general feedback.

References

1. Airsage: Airsage wise technology, <http://www.airsage.com/>
2. LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, J., Sohn, T., Howard, J., Hughes, J., Potter, F., Tabert, J., Powledge, P.S., Borriello, G., Schilit, B.N.: Place lab: Device positioning using radio beacons in the wild. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) PERVASIVE 2005. LNCS, vol. 3468, pp. 116–133. Springer, Heidelberg (2005)
3. Eagle, N., Pentland, A.: Reality mining: sensing complex social systems. Personal Ubiquitous Computing 10(4), 255–268 (2006)
4. Skolnik, J., Chami, R., Walker, M.: Planned Special Events - Economic Role and Congestion Effects. Federal Highway Administration, US-DOT (2008)
5. Alexiadis, V., Jeannotte, K., Chandra, A.: Traffic analysis tools primer, traffic analysis toolbox. Federal Highway Administration, US-DOT (2004)
6. Zhan, B., Monekosso, D.N., Remagnino, P., Velastin, S.A., Xu, L.Q.: Crowd analysis: a survey. Machine Vision and Applications (2008)
7. Kelly, J., Williams, P.W., Schieven, A., Dunn, I.: Toward a destination visitor attendance estimation model: Whistler, british columbia, canada. Journal of Travel Research (2006)
8. Davies, A., Yin, J., Velastin, S.: Crowd monitoring using image processing. Electron. Commun. Eng. J. 7(1) (1995)
9. Ratti, C., Pulselli, R.M., Williams, S., Frenchman, D.: Mobile landscapes: Using location data from cell-phones for urban analysis. Environment and Planning B: Planning and Design 33(5) (2006)
10. Reades, J., Calabrese, F., Sevtsuk, A., Ratti, C.: Cellular census: Explorations in urban data collection. IEEE Pervasive Computing 6(3), 30–38 (2007)
11. Koshak, N., Fouda, A.: Analyzing pedestrian movement in mataf using gps and gis to support space redesign. In: The 9th International Conference on Design and Decision Support Systems in Architecture and Urban Planning (2008)
12. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. Nature 453 (2008)
13. Girardin, F., Vaccari, A., Gerber, A., Ratti, C.: Quantifying urban attractiveness from the distribution and density of digital footprints. Journal of Spatial Data Infrastructures (4) (2009)
14. Ahas, R., Kuusik, A., Tiru, M.: Spatial and temporal variability of tourism loyalty in estonia: Mobile positioning perspective. In: Proceedings of the Nordic Geographers Meeting, NGM 2009 (2009)
15. Social force model for pedestrian dynamics. Phys. Rev. E 51(5) (1995)
16. Bellomo, N.: Modeling Crowds and Swarms: Congested and Panic Flows. In: Modeling Complex Living Systems (2008)
17. Banarjee, S., Grosan, C., Abarha, A.: Emotional ant based modeling of crowd dynamics. In: Seventh International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2005 (2005)
18. Bandini, S., Manzoni, S., Vizzari, G.: Crowd Behaviour Modeling: From Cellular Automata to Multi-Agent Systems. In: Multi-Agent Systems: Simulation and Applications. CRC Press, Boca Raton (2009)
19. Globe, B.: Events and things to do in boston (2009), <http://calendar.boston.com>

20. Gonzalez, M., Hidalgo, C., Barabasi, A.L.: Understanding individual human mobility patterns. *Nature* 453(7196), 779–782 (2008)
21. Calabrese, F., Ratti, C.: Real time rome. *Networks and Communications Studies* 20(3-4), 247–258 (2006)
22. Calabrese, F., Ratti, C., Colonna, M., Lovisolo, P., Parata, D.: A system for real-time monitoring of urban mobility: a case study in rome. *IEEE Transactions on Intelligent Transportation Systems* (2009) (submitted)
23. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)

Indoor Positioning Using GPS Revisited

Mikkel Baun Kjærgaard¹, Henrik Blunck¹, Torben Godsk¹, Thomas Toftkjær¹,
Dan Lund Christensen², and Kaj Grønbæk¹

¹ Department of Computer Science, Aarhus University, Denmark

² Alexandra Institute A/S, Denmark

{mikkelsbk, blunck, tbg, toughcar, dalc, kgronbak}@cs.au.dk

Abstract. It has been considered a fact that GPS performs too poorly inside buildings to provide usable indoor positioning. We analyze results of a measurement campaign to improve on the understanding of indoor GPS reception characteristics. The results show that using state-of-the-art receivers GPS availability is good in many buildings with standard material walls and roofs. The measured root mean squared 2D positioning error was below five meters in wooden buildings and below ten meters in most of the investigated brick and concrete buildings. Lower accuracies, where observed, can be linked to either low signal-to-noise ratios, multipath phenomena or bad satellite constellation geometry. We have also measured the indoor performance of embedded GPS receivers in mobile phones which provided lower availability and accuracy than state-of-the-art ones. Finally, we consider how the GPS performance within a given building is dependent on local properties like close-by building elements and materials, number of walls, number of overlaying stories and surrounding buildings.

1 Introduction

Applying the visions of ubiquitous computing to a variety of domains requires positioning with (*i*) pervasive coverage and (*ii*) independence from local infrastructures. Examples of such domains are fire fighting [1], search and rescue, health care and policing. Furthermore, also many other position-based applications would benefit from positioning technologies that fulfill both requirements [9]. One technology fulfilling (*ii*) is positioning by GPS. However, it has been considered as a fact that GPS positioning does not work indoors and therefore does not fulfill the coverage requirement (*i*). Due to recent technological advances, e.g. high-sensitivity receivers and the promise of an increase in the number of global navigation satellites, this situation is changing.

In 2005, LaMarca *et al.* [11] studied GPS availability with an off-the-shelf receiver for tracking the daily tasks of an immunologist, a home maker and a retail clerk. For the three studied persons, the availability was on average only 4.5% and the average gap between fixes was 105 minutes. To address these shortcomings they proposed fingerprinting-based positioning [8] as a solution. However, for the previously mentioned domains fingerprinting-based solutions are less suitable, given the requirement of fingerprinting collection, the vulnerability to hacking, that e.g. fires might alter the building and the unknown factor of whether or not fingerprinted base stations are taken out, e.g. by a fire.

We have conducted a measurement campaign at several indoor sites, including wooden and brick houses, a public school, a warehouse, a shopping mall and a tower block, to determine to what extent GPS is usable indoors and which performance to expect from it. Furthermore, we intended to link the measured performance to the type of errors affecting GPS as well as to local properties of the buildings like dominating materials and proximity to external walls and windows or surrounding buildings.

In this paper we argue that—when using state-of-the-art receivers GPS—GPS indoor performance is better than suggested in earlier literature. The results of our measurement campaign, which is to our knowledge the most comprehensive of its kind, show good GPS availability in many buildings except for larger ones with thick roofs or walls. The horizontal RMS error in our measurements was below five meters in wooden and below ten meters in most of the brick and concrete buildings investigated. Lower accuracies could be linked to low signal-to-noise ratios, multipath phenomena or bad satellite constellation geometry. We also considered GPS receivers embedded in mobile phones which provided lower availability and accuracy than dedicated receivers.

The rest of this paper is structured as follows: In Section 2 we give a brief introduction and overview of research on GPS and satellite based navigation with a focus on indoor usage. In Section 3 we present our measuring methodology. In Section 4 we present our analysis of the measurement campaign. Finally, Section 5 concludes the paper and provides directions for future work.

2 GPS Primer

GPS satellites send signals for civilian use at the L1 frequency at 1.575 GHz; these signals are modulated with a *Pseudo-Random Noise (PRN)* code unique to each satellite. A GPS receiver tries to *acquire* each GPS satellite's signal by correlating the signal spectrum it receives at L1 with a local copy of the satellite's PRN code. An acquisition is successful, once the local copy is in sync with the received signal, which requires shifting the copy appropriately both in time and in frequency. The latter shift is due to the Doppler effect caused by the satellite's and the user's relative motion. Once a satellite's signal has been acquired, the receiver *tracks* it, that is, the receiver continuously checks the validity of the shift parameters above and updates them if necessary.

Each satellite's signal is modulated not only with its PRN code but additionally with a navigation message, which contains almanac data (for easier acquisition of further satellites) as well as its precise *ephemeris data*, that is the satellite's predicted trajectory as a function of time, allowing GPS receivers to estimate the current position of the satellite. Finally, to achieve precise 3D positioning with a standard GPS receiver via trilateration, the positions of and distances to at least 4 satellites have to be known; those distances can be computed from the time shift maintained while tracking the respective satellites. As a general rule, the more satellites can be tracked, and the wider they are spread over the sky as seen by the user, the more precise the positioning – due to the additional distance data and a satellite geometry resulting in less error-prone lateration.

A popular enhancement of GPS positioning is given by *Assisted GPS (A-GPS)* [17]. A-GPS provides assistance data to GPS receivers via an additional communication channel e.g. a cellular network. This assisting data may consist of e.g. ephemerides

and atmospheric corrections. Also, a cellular network provides means for a rough positioning of the GPS enabled device. A-GPS eases satellite acquisition and can therefore drastically reduce the time to first fix and the initial positioning imprecision of a receiver in *cold start* (i.e. when no initial information about satellite constellations is available): Essentially, A-GPS allows for a *hot start* (precise ephemerides for all satellites available), once the assisting data has been transmitted. Furthermore, A-GPS can improve positioning accuracy by eliminating systemic error sources [12 Chapter 13.4].

GPS performance degrades in terms of both coverage and accuracy when experiencing problematic signal conditions, e.g. in urban canyons and especially in indoor environments. The cause for this is termed signal *fading*, subsuming two fundamental signal processing obstacles: First, when GPS signals penetrate building materials, they are subjected to attenuation, resulting in lower *signal-to-noise ratio (SNR)*. Furthermore, the signal is subject to *multipath phenomena*: Reflection and refraction of the signal results in multiple echoes of the *line-of-sight (LOS)* signal arriving at the receiver. Low signal-to-noise ratios and multipath handicap both acquiring and tracking GPS signals and usually result in less reliable positioning due to less suitable satellite geometry and individual time shifts measurements being less accurate. *High-Sensitivity GPS (HSGPS)* [10] receivers are specifically designed for difficult signal conditions, i.e. to alleviate the above problems. HSGPS is claimed to allow tracking for received GPS signal strengths down to -190 dBW: three orders of magnitude less than to be expected in open-sky conditions [12]. These thresholds are constantly being improved using new processing techniques [17 Ch. 6]. Note, that for acquiring signals at cold start, a somewhat (around 15dBW) higher signal strength is usually necessary, as during acquisition reliable time and frequency shifts of the signal have not only to be maintained, but instead searched for in a wide spectrum.

With respect to future improvements towards satellite based indoor positioning, note also that the upcoming Galileo system is a *Global Navigation Satellite Systems(GNSS)*, like GPS, and will soon be interoperable with the latter, resulting in roughly a doubling of GNSS satellites available [12 Ch. 3]. Combined satellite constellations will yield, in effect, better geometries at the user position, improving positioning accuracy, especially indoors, where signals from only parts of the sky may be available. Other upcoming improvements for indoor GNSS are provided by the modernized public signal structures of GPS and Galileo, allowing improved tracking of weakened signals via pilot channels, and yielding additional protection against multipath-induced inaccuracies [2]. In the GNSS community indoor positioning and respective obstacles and improvements are being investigated, see, e.g., Teuber *et al.* [16], Paonni *et al.* [13], Lachapelle *et al.* [10], Watson *et al.* [18] and references therein. This paper adds to this line of work but from an application-oriented perspective using real-world measurements to investigate where and to what extent one can employ GPS for indoor positioning.

3 Measurement Campaign Methodology

In the following, we describe the measurement campaign, the equipment used, as well as methodology regarding measurement collection procedures and choice of in-building locations; for a more thorough justification of the chosen methodology see also [3].

3.1 Receiver Equipment Employed

Throughout our campaign, we employed a u-blox LEA-5H Evaluation Kit and a SiRF-Star III BU-353 USB GPS receiver as examples of dedicated receivers, and a Nokia N95 8GB driven by Texas Instruments' Navilink 4.0 GPS5300 chip as an example for a currently used in-phone GPS receiver system. The u-blox receiver is specified to have a -190dBW (-175dBW) threshold for tracking (respectively, acquisition) and was connected to a 48x40x13mm u-blox ANN-MS patch antenna, providing 27dB gain and specified with a 1.5dB noise figure. We will focus on the measurements obtained with this dedicated receiver and note in passing, that the SiRF product performed equivalently, though slightly poorer which might be solely due to the u-blox's high quality external patch antenna. To obtain A-GPS assistance data [17], we connected the u-blox receiver to a N95 phone.

The two classes of receivers considered differ not only in performance but also in price, energy consumption and size: Whereas the larger and more power consuming dedicated receivers will be used in specific scenarios mentioned in the introduction, such as search and rescue operations, the Nokia N95 in-phone receiver represents typical hardware for the every-day consumer use of location based services. Furthermore, given the pace of development the chosen dedicated receiver allows an outlook on the performance of future in-phone GPS receivers used, e.g., for location-based services¹.

3.2 Data Collection Procedures

During our campaign, we focused on static measurements at a number of locations per building, partially in order to eliminate effects from the receiver's recent history of measurements. Consequently, we reset the receivers prior to each measurement, thereby minimizing the effects of, e.g. Kalman filtering techniques, which exploit recent measurement history and therefore potentially pollute static measurements w.r.t. both locations and durations, see, e.g., Brown and Hwang [4].

To focus on and to fairly compare signal conditions at individual indoor locations, we also decided against an on-person receiver setup. Instead, each of the GPS enabled devices was mounted on the seat of a light-weight wooden chair, spaced 20 cm apart from the other receiver devices to remove the chance of any near-field interference. Note though, that the measurements carried out within the shopping mall, the warehouse and the tower block were conducted during business hours, providing realistic pedestrian traffic conditions, see also [14] for the impact of pedestrian traffic on GPS performance. The chair was then placed at the in-building measurement locations chosen and we collected GPS measurements using the following procedure:

After initializing the programs for logging GPS data at 1 Hz, we "hot started" the Nokia and the u-blox receivers. Note, that by hot start we refer to a initialization using A-GPS data. If the hot start of the u-blox receiver was successful, i.e., if it within 10 minutes produced a position fix, we subsequently logged the u-blox receivers' NMEA formatted data for 5 minutes; then we repeated as above, but this time cold starting the receivers without A-GPS. In case the hot start of the u-blox was not successful, we

¹ The generation 6 of the u-blox receiver has been improved over the version used here, specifically focusing on low energy consumption, allowing for more economic use in mobile gadgets.

produced a successful hot start at a nearby location and walked back to the measurement location. If the receiver still produced position fixes upon arrival, we logged the receivers' data for 5 minutes.

3.3 Choosing Measurement Locations

For choosing where to measure GPS reception within the buildings, we overlaid the larger buildings' floor plans with a regular grid, choosing measurement locations as the centers of the grid cells, where feasible, i.e. where these centers fall within the respective building. We chose this strategy to avoid biases induced by alternative approaches in which locations are picked so to reflect environmental conditions which are—*by a priori* hypotheses—associated with specific GPS signal reception conditions.

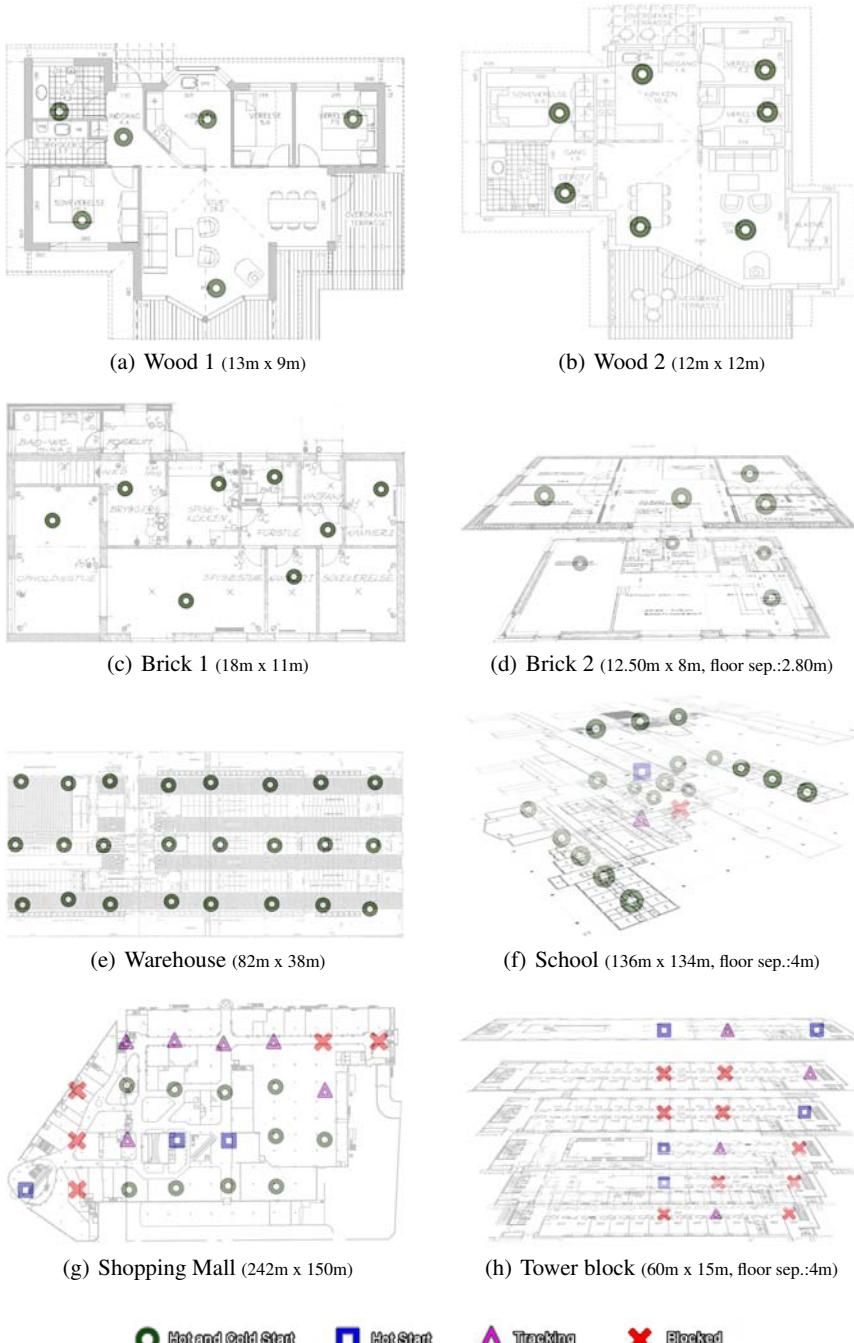
4 Results

Using the measurement procedures described above allows us to explore the performance of GPS positioning in various environments using state-of-the-art receiver technology. In section 4.1 we will introduce the environments investigated in our measurement campaign and characterize them with regards to availability, the most fundamental measure for GPS performance. Subsequently, we will cover further GPS performance measures, namely time-to-first-fix in Section 4.2 and positioning accuracy in Section 4.3. Throughout this section, we not only present performance measures for the individual measurement locations, but also elaborate on general rules which govern GPS indoor performance, as observable in the analysis of our measurement campaign. In Section 4.4 we give results for measurements using the Nokia N95 in-phone receiver, comparing them to the measurements obtained from dedicated receivers.

4.1 GPS Availability and Signal Strength

The GPS availability results for the eight environments chosen for our campaign are illustrated in Figure 1. The figure shows for each in-building measurement location whether and by which means we were able to acquire GPS fixes. As described by the figure legend, we categorize availability performance into 4 categories: (i) both *hot start* and *cold start* (indicated by a green circle) were successful; (ii) a *hot start* but no *cold start* (indicated by a blue square) was successful; (iii) neither hot or cold starts were successful, but *tracking* was, that is acquiring a GPS position at a location with high GPS availability and moving to the measurement location where GPS upon arrival continued to produce position fixes (indicated by a purple triangle); (iv) and finally *blocked* where no GPS position fixes could be established (indicated by a red cross).

A main factor impacting to which extent and quality one can get GPS fixes at specific in-building locations are the surrounding building structures and elements. Therefore, complementing Figure 1, we listed in Table 1 for the environments investigated the respective dominating materials used for external and internal building elements. The table also contains approximations, compiled from various sources [6][15][19], for



Hot and Cold Start

Hot Start

Tracking

Blocked

Note, that the distances between floors are modified in the figure in order for all grid points to become visible.

Fig. 1. Overview of GPS availability in various building types

Table 1. Building materials and their attenuation properties in the buildings investigated

Building Type	Walls			Roof	dB
	External	dB	Internal		
Wood 1	Wood	2.40	Wood	2.40	Tiles 5.19
Wood 2	Wood	2.40	Wood	2.40	Tiles 5.19
Brick 1	Double Brick	10.38	Brick	5.19	Fiber Cement N/A
Brick 2	Double Brick	10.38	Brick	5.19	Tiles 5.19
School main building + right wing annex + left wing	Double Brick Brick and Concrete	10.38 14.76	Brick Concrete	5.19 9.57	Tiles Tiles 5.19 5.19
Warehouse	Fiber Cement and Curtains/Openings	N/A N/A	Equipment	N/A	Fiber Cement N/A
Shopping Mall	Reinforced Concrete Tinted Glass Glass	16.70 24.44 2.43	Brick	5.19	Flagstone Sand Felt roofing Concrete N/A 2 9.57
Tower block	Double Brick around Concrete	19.95	Brick	5.19	Tiles 5.19

the attenuation (w.r.t. GPS L1 frequency signals and assuming an incident angle of 0 degree) caused by the respective building materials (assuming common respective thicknesses).² The attenuation values listed directly impact the signal-to-noise-ratio of GPS signals indoors, since, as a rule of thumb, a penetrated material's attenuation value is to be subtracted from the received signal-to-noise-ratio as experienced outside the given building. For signals penetrating multiple layers of building materials, attenuation can be considered at least additive. The average signal-to-noise-ratio over time for the 4 satellites with the strongest received SNR is shown in Figure 2 for all measurements, grouped by building, see, e.g., Misra and Enge [12] for relating signal-to-noise-ratios and GPS signal power. Differences in the SNR figures within one building are naturally due to properties of the individual in-building locations. Such properties include the number and distances to building elements such as walls and roofs: For example, Teuber *et al.* [16] concluded from their measurements, that the power of received GNSS signals does not only depend on the building materials penetrated, but also further decreases with the distance traveled after the respective penetrations. In general, our measurements confirmed that observation.

Per-case availability analysis. When looking at Figure 1 the GPS availability seems promising for both the two wooden houses 1(b) and 1(a) as well as for the two brick houses 1(c) and 1(d), of which only the last one is a multi-story building. However, in the other buildings GPS is only partially available, and in order to understand these variations we will go deeper into the analysis of these particular buildings.

The larger 82m x 38m warehouse 1(e) is a relatively open environment, with just cloth curtains between roof and lowered outer walls, and four 50cm wide skylights

² For GNSS frequencies lower than L1, e.g., L2 and L5, the attenuation for most of the listed materials is somewhat lower, see, e.g., [6, Table 3]. For further studies on the strength of GPS frequency signals in indoor environments see also [5, Ch. 9.4.2] and references therein.

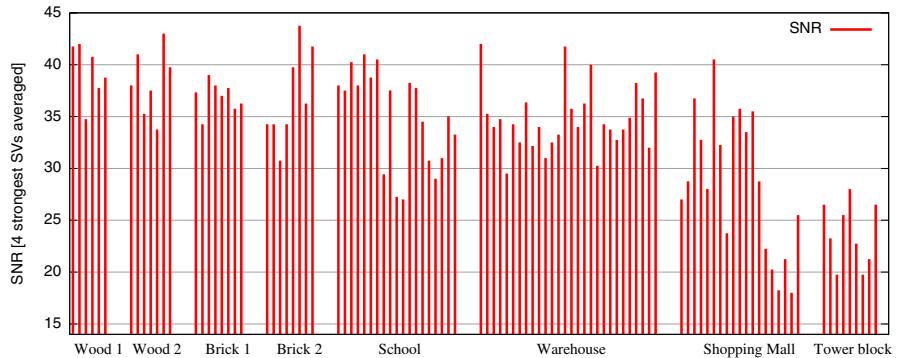


Fig. 2. Averaged signal-to-noise ratio at individual measurement locations

along the roof. Consequently, it allows for proper GPS fixes. Further analysis shows that reception was difficult and SNR low for signals which had to penetrate interior elements of the warehouse—whereas signals entering close to or through skylights were received much stronger. That GPS signals are received strongly only from certain parts of the sky is observable in the skyplots [7] in Figure 3 for two exemplary measurement locations close to exterior walls of the warehouse. In the skyplots, the 3 concentric circles represent 0, 30, and 60 degrees elevation, respectively. Depicted are the location of the satellites, tracked by the receiver during the respective hot start measurement period. Individual satellites are identified by the id of the PRN code, they are sending, respectively. The individual positions over time of each shown satellite are depicted by “+” symbols, where the symbol’s color indicates signal-to-noise ratio, as experienced by the u-blox receiver at the respective measurement location and according to the color scale given in the figure. A green arrow trailing the orbit of a satellite signals its direction of movement. The pseudorange error for each satellite and for each individual time instance of reception during the measurement period are sketched in blue color, according to the scale given in the figure, and—for presentation purposes—perpendicular to the respective satellite’s direction of movement.³

The school building at Figure 1(f) forms an H with two single-story wings and one three-story middle section and finally a single-story annex. First, due to the skylight windows signals have easy access to locations in the two wings. In the annex and also in the middle section on the second floor strong signals were present. Second, the first floor allowed for receivable but weaker signals, in particular at the location at the center of the middle section. Here a cold start was not successful, possibly due to the attenuation caused by the top floor, and due to the location being in a wide part of the building. Third, in the basement the signals are attenuated by the two top floors and only due to the relatively open area at the center were we able to track the signal there. At another basement location no GPS fix were achieved at all.

³ Note that while a GPS receiver can only output pseudorange errors w.r.t. the estimated position, they can be transformed into pseudorange errors w.r.t. the actual receiver position, given properly surveyed ground truth by means of, e.g., satellite imagery, building floor plans and laser ranging, as done by the authors, see also [3, Ch.3].

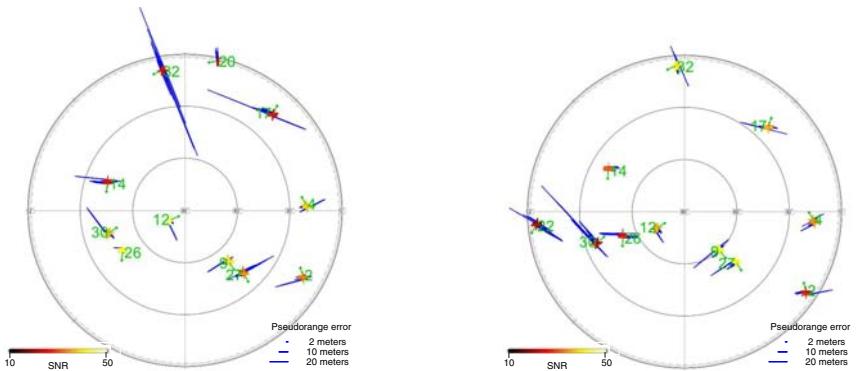


Fig. 3. Skyplots of measurements at two locations inside the warehouse

Figure 1(g) depicts the middle floor a three-story shopping mall. The grid points will be referred to by coordinate tuples, where the bottom left corner would be (1,4), and the upper right would be (8,1), which is consistent with Figure 9, which will be discussed in Section 4.3 and which depicts most of the mall's middle floor plan, overlaid with skyplots. The center (3-5,3) of the middle floor is covered by a top floor of smaller size, and some other locations (2,2-4), (3-8,1), (7,1-3) are not only covered by a roof, but also by a parking deck. As a first observation, the grid points where both a cold and a hot start was possible are located in the single-story part of the building where only a top felt roof attenuates the GPS signals, with the exception of one grid point at the bottom of the second most right row (7,3), which is covered by the parking deck. Second, the signal could at least be tracked at all grid points that are covered by the second story, which consists of shops, offices and an atrium-like opening from the top floor roof down to the basement. Third, hot starts could also be performed at the grid points (4,3) and (5,3)—most likely due to the closeness to the glass atrium covering all floors. This hypothesis is supported by the relatively high SNR values experienced at these two locations (as compared with other measurement locations beneath the same heavy roof structure). Fourth, grid point (1,4), located in another atrium, presumably depicts the highly attenuating effect of tinted glass, see Table I, resulting in comparatively low SNR values and allowing only for a hot, but not for a cold start. Fourth, all 5 blocked locations are among the 11 grid points covered by the top floor parking lot, where the separation to the middle floor consists of layers of steel-reinforced concrete, sand and flagstones. With the exception of point (7,3) only tracking is possible at the remaining 6 points, implying that GPS reception is still difficult. Note also, that, though not depicted in Figure 1(g), measurements were performed also in the basement, which provided position fixes only near exits and near the atriums spanning all floors.

Figure 1(h) shows all stories except the ground floor of a seven-story office building built in a tower block fashion. Averaged over the in-building measurement locations, this building showed the highest signal attenuations experienced in any of the explored buildings. This is only partially due to the outer double-brick and concrete walls. Additionally, our measurements showed that SNR was significantly impacted by the inner walls or ceilings, the respective signals had to penetrate. Fittingly, all blocked locations,

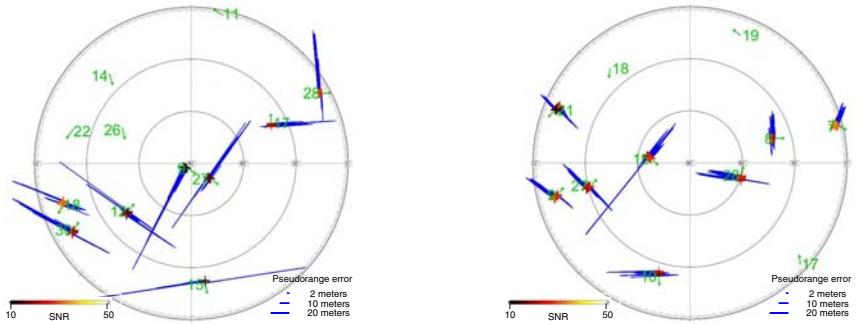


Fig. 4. Skyplots of measurements on 2nd (left) and 7th (right) floor of a tower block

apart from those residing on the staircase, are in narrow aisles surrounded by small offices. Furthermore, the two measurement locations at the 2nd and 3rd floor where hot starts were possible are both located adjacent to a two story library which means that GPS signals from lower elevation satellites have to penetrate only the external walls. Similarly, reception was possible at all three locations on the top floor, where signals can pass with less attenuation, not being obstructed by multiple inner building elements.

The poor reception on the lower floors can also be attributed to an additional shielding effect caused by two adjacent four-story buildings. These are placed in the same major direction as the building depicted and in immediate continuation and on opposite sides of what is depicted as the rightmost end of the building. The signal attenuation by these two buildings are contributing to the GPS unavailability on the second, third and fourth floor in the staircase to the right, additional to the attenuation caused by the concrete-built staircase as well as building elements further above. Consistent with this explanation, on the same staircase, but on the highest three level (clear from the shielding buildings), tracking, and for the 5th and 7th even hot starts, were successful.

To illustrate exemplary the effects of attenuation by multiple building stories, Figure 4 shows skyplots [7] for the middle location in Figure 1(h) on the 2nd and 7th floor, respectively. The observed SNR values are generally, and especially around the zenith, higher at the location on the highest floor, since the latter is separated from open sky by less elements of its own (as well as of the neighboring) building(s). Noteworthy, the low SNR values around the zenith are typical for the measurements we carried out at lower levels of multi-story buildings, but stand in contrast to the outdoor situation where SNR values generally increase with the satellites' elevation w.r.t. the receiver position. Both skyplots depict also the positive effect of the tower block's windows in east and west direction, adding to the proper signal reception from low elevation satellites.

Summary. We have found that both signal-to-noise-ratios as well as, in result, the availability of GPS indoors—using today’s receiver technology—is generally more promising than suggested by earlier positioning literature. Furthermore, covering many different building types, we found, mostly in confirmation with empirical studies for different individual environments, that GPS availability is negatively impacted by: the number of overlaying stories, the roof material, e.g. reinforced steel, in contrast to more

favorable materials such as felt roofing or fiber-cement, as well as wall materials and the number of walls, the distance to the walls separating the receiver from the outside and the closeness to surrounding buildings.

4.2 Time To First Fix

The time to first position fix is prolonged, where acquisition of weaker and refracted signals is necessary, in particular indoors. Figure 5 plots on a logarithmic scale time to first fix in seconds for hot and cold starts of the u-blox receiver in three environments.

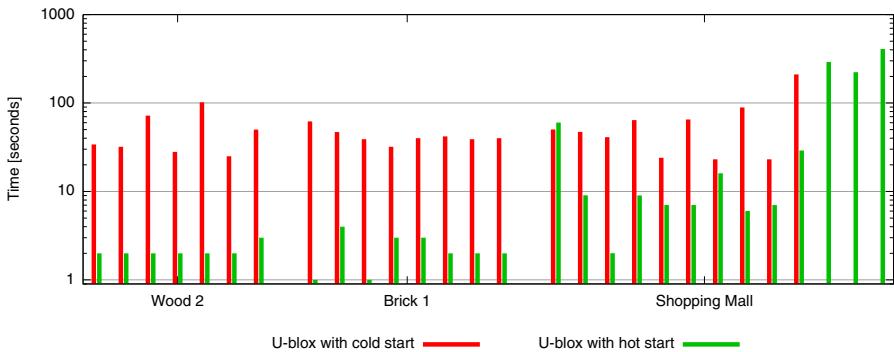


Fig. 5. Time to first fix for three measurement sites

Inside both buildings ‘wood 2’ and ‘brick 1’ the u-blox receiver shows a fast hot acquisition of less than four seconds. As expected, the cold starts take longer, around 40 seconds on average, with some faster at half a minute and some slower. In comparison, for outdoor use the technical specification of the u-blox receiver states that hot starts take less than one second and cold starts take about 29 seconds. In the shopping mall the hot starts take around ten seconds on average but at three locations the time increases to several minutes. Comparing SNR and time to first fix for each measurement location, one can observe a strong dependency between the weakness of signals and the time it takes to acquire the signals. This implies that in locations with weak signals one can expect high values for time to first fix. For cold starts the average time-to-first-fix is around 60 seconds but at the three last locations cold start were not even possible within the 10 minute limit.

4.3 Accuracy

To study the GPS positioning accuracy in the environments investigated using a dedicated receiver we have compared the u-blox receiver’s GPS position fixes with manually surveyed ground truth positions. Figure 6 shows for each measurement location, which allowed for a hot start or tracking the root mean squared 2D and 3D positioning errors, averaged over the five minutes of data gathered. Figure 7 shows per building the cumulative 2D error distribution, averaged over all position fixes gathered at in-building

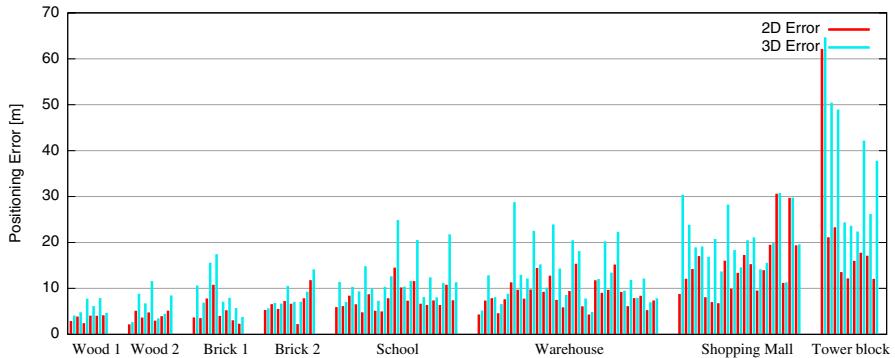


Fig. 6. RMS positioning error in 2 and 3 dimensions

measurement locations. These CDFs yield popular comparative measures of GPS quality, implying an order on the investigated buildings—which is noticeably invariant for most of the popular confidence intervals, e.g., 50th, 67th (i.e. RMS), and 95th percentile. Note, that u-blox claims to allow for a horizontal RMS position error of 3m or less, a claim which holds for outdoor measurements we carried out in open-sky surroundings. The accuracy achieved for cold starts (not shown) averaged over 5 minutes is as expected lower due to the usually small number of satellite ephemerides known and the resulting poor DOP values, when achieving first position fixes without assistance data. The accuracy, though, usually converges over time to the one for hot starts, as more parts of the almanac and precise ephemerides for newly acquired satellites can be decoded. For the remainder, when referring to or visualizing measurement details, we implicitly refer to ‘hot start’ measurements where successful and to ‘tracking’ ones, where not.

Similar to the performances measures of availability and time to first fix, also accuracy is impaired by signal attenuation. This is mainly due to the following two reasons: First, signal attenuation may lead to fewer satellites being tracked and therefore to less favorable satellite constellation geometries. Second, low signal-to-noise-ratio of

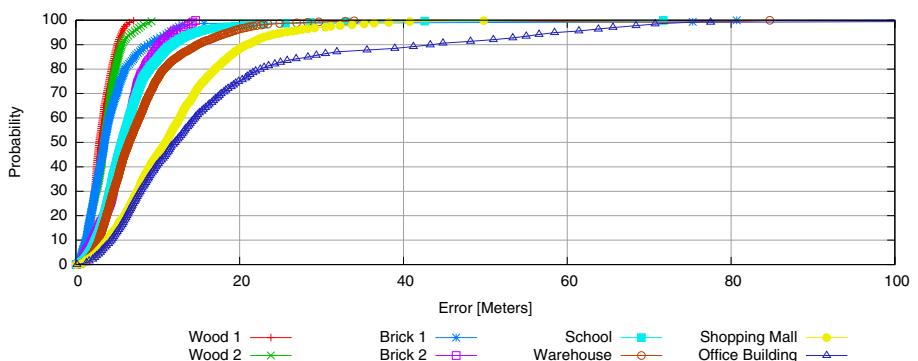


Fig. 7. Cumulative distribution functions (CDFs) of 2D positioning errors per building

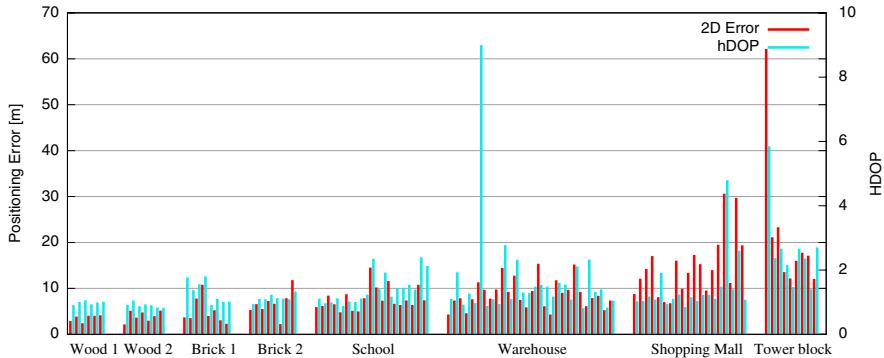


Fig. 8. RMS 2D positioning error and respective horizontal DOP

received signals may result in less precise tracking of the signal, and therefore in less accurate measurements of the distance to the satellite. Figure 8 shows both the RMS 2D error as well as the horizontal DOP value, as an indicator for the negative impact of the satellite geometry with values below 1 being considered ideal. In an outdoor setting, a linear dependency of DOP value and positioning inaccuracy should be observable, and so it is also, to a large extent, in our measurements. Deviations from this linear dependency are mostly constituted by the second main reason for GPS inaccuracy indoors: multipath phenomena. In the following, we will discuss both positioning accuracy and the impact of multipath phenomena in a per-environment analysis, relying foremost on analysing skyplot visualizations yielding SNR and the pseudorange errors for individual satellites received at the measurement locations. Note, that within a GPS receiver the measured pseudoranges are the main basis for the subsequent positioning computation via lateration.

For the wooden buildings (see Figure 1(a) and Figure 1(b), respectively) positioning accuracy is close to outdoor levels, at 3.6m and 4.0m, respectively, for the average over the RMS 2D positioning errors shown in Figure 6. The good accuracy is partially due to the low attenuation of wooden building materials. Additionally, the small size of the building means closeness to outer walls and few or none obstructing inner building elements. Finally, multipath phenomena are weak, because the line-of-sight signal is as expected strong, and because the time-of-arrival difference between line-of-sight signal and echoes is small, since reflecting buildings elements are generally close by. For the two brick houses the positioning errors were on average, at 5.0m and 6.7m, respectively, only slightly larger than the smaller houses built from less attenuating wood. Together with the results from the recent sections, this suggests that modern dedicated receivers can cope well with the indoor challenges within small houses.

Within the school depicted in Figure 1(f) the 2D positioning errors are significantly larger for locations buried deep inside the building, i.e., in the center and especially in the basement of the building, increasing the average error to 7.8m, whereas for other locations the error is on the same level as for the brick houses.

The measurements in the warehouse (Figure 1(e)) showed an average 2D error of 8.8m and the average HDOP of 1.7 was considerably higher than for the buildings



Fig. 9. Measurement locations in the shopping mall, overlaid with skyplots showing SNR and pseudorange error, as observed by the u-blox receiver

mentioned above, suggesting that the large window and skylight areas allow for easy access for GPS signals, but only from certain parts of the sky, as visible in Figure 3.

Within the main floor of the shopping mall, depicted in Figure 1(g), the RMS positioning errors, averaging at 14.8m in the plane, deviated more than in any of the building mentioned above; this observation correlates with the heterogeneity of both architecture and building materials used in the mall. To support the discussion of the different signal conditions and resulting accuracies observable, Figure 9 shows skyplots, as introduced in Section 4.1, for all measurement locations in the center part of the mall's main floor. All locations in the top row and rightmost column lie beneath the mall's parking lot—causing low SNR values. Interestingly, the pseudorange errors for satellites around the zenith are not necessarily large. Notably, though, location (7,2) shows different data: Mostly in the sky part below which windows lie, reflections through these windows seem to be stronger than the line-of-sight signals, resulting in large multipath-induced errors.⁴ The biased pseudorange measurements lead to strongly biased positioning,

⁴ Note, that multipath-induced errors can be con- or destructive: Depending on the relative phase of the incoming signal versions, they either lengthen or shorten the pseudorange measured.

resulting in the largest horizontal RMS error of all locations in the mall, except (4,1). At the latter location tracking was possible only for 4 satellites and for a short amount of time, leading to the within the mall by far largest horizontal DOP values of over 4.

Another area where accuracy is strongly impacted by multipath phenomena is under the atrium roof. The atrium located between locations (4,3) and (5,3) spans all three roof and provides signal echoes easy access especially to locations (3-5,3) which are otherwise covered by the mall's top floor. Consequently, the skyplots for (3,3) and (5,3) suffer from large pseudorange errors indicating that the echoes hinder precise tracking of the line-of-sight (versions of the) GPS signals, resulting in biased position fixes, deviating in particular directions from the true location. Such an effect does not occur and the pseudorange error is small, in case a satellite is received in direct line of sight through the atrium, as e.g., PRN 23 as received from (4,3).

When rerunning measurements at day-times yielding considerably different satellite constellations, we noted only minor changes in GPS performance measures. Exceptions occurred where SNR is rather good, but multipath phenomena impact the positioning strongly, depending on the current satellite constellation: Of all mall locations investigated, location (4,3) showed the largest deviation in 2D error, from an original 8.1m RMS to 17.2m for the rerun of the measurement, averaged over 5 minutes, respectively.

The tower block (Figure 1(h)) exhibits, averaging over measurement durations and locations, the by far poorest SNR (of 23), and the highest HDOP (of 2.7) and, consequently, also the largest horizontal errors (of 21.7m RMS) amongst all investigated environments. While the highest floor shows acceptable reception, SNR and pseudorange errors are worse on lower floors as shown for the 2nd floor in Figure 4. Consequently and consistent with the results obtained in the other investigated buildings, the attenuated and indirect signals yield here a much larger HDOP value of 5.8 and horizontal RMS error of 62.1m than on the top floor with 1.5 and 12.2m, respectively.

Summary. The accuracy in the four wooden and brick houses investigated was good using the dedicated u-blox receiver due to it being separated from the outside both by only few building elements and also only short distances, resulting in low signal attenuation and dispersion and in only small delays of multipath echoes. If more building elements get between the receiver and the outside, as in the basement of a school or a mall, under a roof parking lot or deep inside a tall building, signal attenuation impacts both availability and positioning accuracy, since only few satellites and only in restricted parts of the sky can be acquired leading to poor satellite constellation geometries.

Especially in environments of heterogeneous architecture like in the investigated mall, GPS accuracy varies considerably and can be strongly biased and impaired by multipath phenomena, e.g., when window areas allow for echoes being potentially stronger than the line-of-sight signal which may have to penetrate strongly attenuating building elements to be received directly.

4.4 Embedded GPS Receivers

Embedded GPS receivers in mobile phones are restricted both in terms of power consumption and antenna type and size. In result, such receivers are less sensitive than those

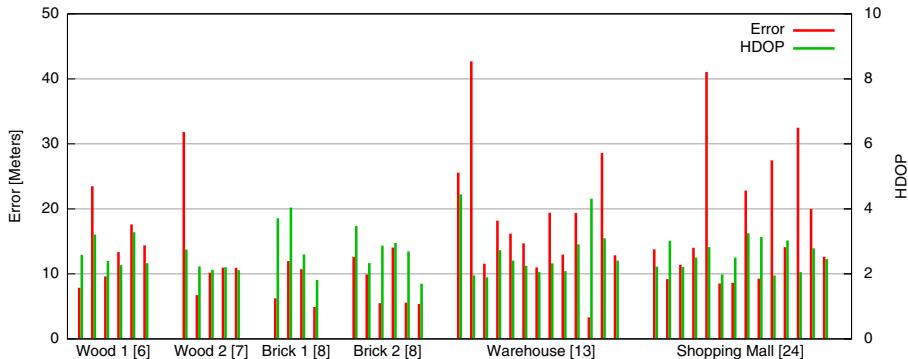


Fig. 10. Horizontal RMS errors and DOP values, using the N95 in-phone receiver

typically used for standalone receivers, implying less sensitive antennas and weaker amplification stages, see also [5 Ch. 9.4.2.1]. Therefore, it is relevant to consider how well an embedded GPS receiver performs compared to the state-of-the-art receiver we relied on in the previous sections. As mentioned in Section 3, we collected data with a Nokia N95 8GB phone which employs a Texas Instruments GPS chip launched in 2006.

Figure 10 shows the horizontal RMS error and the average horizontal DOP value for each measurement location in the six environments where we collected measurements using the N95 embedded receiver. The labellings at the bottom of the figure include the number of measurement locations for each environment. By comparing for each environment the latter number to the number of bars shown, one can comment on the availability for the N95. Generally, the N95 allows for GPS positioning in fewer locations than the u-blox receiver, except for the house ‘wood 1’ and the warehouse, where availability was equivalent.

The horizontal RMS errors measured in the four houses average around 10 meters, for the warehouse and the shopping mall even higher. Particularly for the four houses,

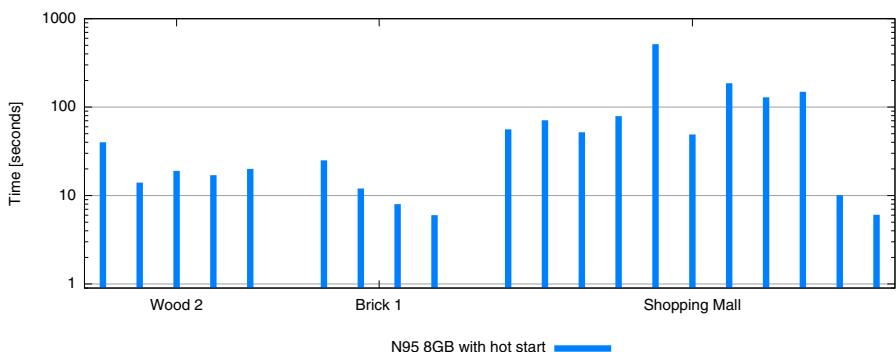


Fig. 11. Time to first fix for the N95 for hot starts

the RMS values are twice as bad as when using the u-blox receiver. Due to the N95 acquiring consistently fewer satellites than the more sensitive u-blox receiver, the HDOP results for the N95 are usually higher than 2 and twice as large as for the u-blox, which yields the main explanation for the lower positioning accuracy of the N95. The time to first fix for hot starts is shown on a logarithmic scale in Figure 11. For the ‘wood 2’ and ‘brick 1’ environments it is around 10 seconds, thus longer than the 1-3 seconds used by the u-blox receiver. For the shopping mall it averages around 90 seconds—much more than the on average 10 seconds used by the u-blox.

In summary, the embedded N95 GPS receiver is able to provide positioning in considerably fewer indoor environments than the u-blox receiver. Furthermore, the time to first fix is considerably longer and positioning errors are twice as large.

5 Conclusions

In this paper we improve on the understanding of indoor GPS reception characteristics by analyzing results from a measurement campaign covering eight different buildings. We have found that both signal-to-noise-ratios as well as, in result, the availability of GPS indoors using state-of-the-art receiver technology is generally more promising than suggested in the positioning literature. Furthermore, covering many different building types, we found that GPS availability is negatively impacted by: the number of overlaying stories, the roof material, as well as wall materials and the number of walls and the closeness to surrounding buildings. Time to first fix with at hot starts, i.e. using A-GPS, generally took less than ten seconds. However, at some locations longer time was required, occasionally more than a minute. Especially for battery-powered devices this might be a drawback as longer time to first fix will consume extra power. The 2D root mean squared accuracy of the measurements was below 5 meters in the wooden and below 10 meters in most of the brick and concrete buildings. Low accuracies can be linked, depending on the environment’s characteristics, to either low signal-to-noise ratios, multipath phenomena or poor satellite constellation geometries. We also carried out measurements using GPS receivers, embedded in mobile phones, which provided considerably lower availability, lower accuracy and longer time to first fix than the state-of-the-art receivers employed in the campaign.

Our results indicate for the application domains mentioned in the paper, that GPS can be used as a positioning technology to provide situational awareness at a building-part granularity, especially when A-GPS is available, yielding an accuracy level of tens of meters and a time to first fix in the range from a few seconds to minutes. GPS though, does not currently provide instantly available indoor positioning accurate to the meter as might be crucial for some indoor applications, e.g. fire fighters navigating in burning buildings. Therefore, another use of indoor GPS would be as a complementary indoor positioning technology, e.g., to help fighting the error growth over time of inertial positioning systems when available. The results are also indicative for the performance of future embedded GPS devices, as state-of-the-art receivers are being constantly miniaturized and power optimized.

The line of work presented in this paper naturally gives rise to further items of research, potentially inspiring or giving input to improved GPS positioning algorithms. It

would be relevant to study systematically to which extents in-building position parameters like distance and angle to the closest wall, window or room corner affect signal strength and quality as measured at the receiver position. Whereas such dependencies were found and formulated already by Teuber *et al.* using data gathered in a single office building, a validation of such dependencies in other real-world indoor environments is yet to be done. Furthermore, it would be relevant to conduct new measurement campaigns to evaluate the impact on indoor GNSS performance of the new GPS and Galileo signals and satellites as they become available.

Acknowledgements. We thank Anders and Jens Emil Kristensen for their help in collecting measurements and acknowledge the financial support granted by the Danish National Advanced Technology Foundation for the project "Galileo: A Platform for Pervasive Positioning" (009-2007-2) and by the Danish National Research Foundation for MADALGO - Center for Massive Data Algorithmics.

References

1. Angermann, M., Khider, M., Robertson, P.: Towards operational systems for continuous navigation of rescue teams. In: Proc. Position, Location and Navigation Symposium (2008)
2. Avila-Rodriguez, J.-A., Hein, G., Wallner, S., Issler, J.-L., Ries, L., Lestarquit, L., de Latour, A., Godet, J., Bastide, F., Pratt, T., Owen, J.: The mboc modulation. a final touch for the galileo frequency and signal plan. Inside GNSS 2(6), 43–58 (2007)
3. Blunck, H., Kjærgaard, M.B., Godsk, T., Toftkjaer, T., Christensen, D.L., Grønbæk, K.: Empirical analysis and characterization of indoor gps signal fading and multipath conditions. In: Proc. 22nd Intl. Techn. Meeting Satellite Division Inst. of Navigation, ION GNSS (2009)
4. Brown, R., Hwang, P.Y.C.: Introduction to random signals and applied Kalman filtering: with MATLAB exercises and solutions, 3rd edn. (1997)
5. Bullock, J., Floss, M., Geier, G., King, M.: Integration of gps with other sensors and network assistance. In: Kaplan, E.D., Hegarty, C. (eds.) Understanding GPS: Principles and Applications, ch. 9, 2nd edn. Artech House, Reading (2006)
6. Hein, G., Teuber, A., Thierfelder, H., Wolf, A.: Fighting the fading - part 2. Inside GNSS (2008)
7. Hill, S.: Plotting pseudorange multipath with respect to satellite azimuth and elevation. GPS Solutions 8(1) (2004)
8. Kjærgaard, M.B.: A Taxonomy for Radio Location Fingerprinting. In: Proceedings of the Third International Symposium on Location and Context Awareness (2007)
9. Küpper, A.: Location-Based Services: Fundamentals and Operation, October 2005. Wiley, Chichester (2005)
10. Lachapelle, G., Kuusniemi, H., Dao, D., MacGougan, G., Cannon, M.: HSGPS signal analysis and performance under various indoor conditions. Navigation, Inst. of Navigation 51(1), 29–43 (2004)
11. LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, J., Sohn, T., Howard, J., Hughes, J., Potter, F., Tabert, J., Powledge, P., Borriello, G., Schilit, B.: Place Lab: Device Positioning Using Radio Beacons in the Wild. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) PERVASIVE 2005. LNCS, vol. 3468, pp. 116–133. Springer, Heidelberg (2005)
12. Misra, P., Enge, P.: Global Positioning System: Signals, Measurements, and Performance. In: Navtech, 2nd edn. (2006)

13. Paonni, M., Kropp, V., Teuber, A., Hein, G.: A new statistical model of the indoor propagation channel for satellite navigation. In: Proc. 21st Intl. Techn. Meeting Satellite Division Inst. of Navigation, ION GNSS (2008)
14. Sokolova, N., Forsell, B.: Moderate pedestrian traffic: Indoor hsgps receiver performance. European Journal of Navigation 5(3), 2–7 (2007)
15. Stone, W.: Electromagnetic signal attenuation in construction materials. NIST Construction Automation Program Report No. 3, National Inst. Standards and Technology, U.S. (1997)
16. Teuber, A., Paonni, M., Kropp, V., Hein, G.: Galileo signal fading in an indoor environment. In: Proc. 21st Intl. Techn. Meeting Satellite Division Inst. of Navigation, ION GNSS (2008)
17. van Diggelen, F.: A-GPS: Assisted GPS, GNSS, and SBAS. Artech House (2009)
18. Watson, R., Lachapelle, G., Klukas, R., Turunen, S., Pietil, S., Halivaara, I.: Investigating gps signals indoors with extreme high-sensitivity detection techniques. Navigation, Inst. of Navigation 52(4), 199–213 (2006)
19. Williams, K., Greeley, R.: Radar attenuation by sand: laboratory measurements of radar transmission. IEEE Transactions Geoscience and Remote Sensing 39(11), 2521–2526 (2001)

Specification and Verification of Complex Location Events with Panoramic

Evan Welbourne, Magdalena Balazinska, Gaetano Borriello, and James Fogarty

Computer Science & Engineering

University of Washington

Seattle, WA 98195 USA

{evan,magda,gaetano,jfogarty}@cs.washington.edu

Abstract. We present the design and evaluation of Panoramic, a tool that enables end-users to specify and verify an important family of complex location events. Our approach aims to reduce or eliminate critical barriers to deployment of emerging location-aware business activity monitoring applications in domains like hospitals and office buildings. Panoramic does not require users to write code, understand complex models, perform elaborate demonstrations, generate test location traces, or blindly trust deterministic events. Instead, it allows end-users to specify and edit complex events with a visual language that embodies natural concepts of space and time. It also takes a novel approach to verification, in which events are extracted from historical sensor data traces and then presented with intelligible, hierarchical visualizations that represent uncertainty with probabilities. We build on our existing software for specifying and detecting events while enhancing it in non-trivial ways to facilitate event specification and verification. Our design is guided by a formative study with 12 non-programmers. We also use location traces from a building-scale radio frequency identification (RFID) deployment in a qualitative evaluation of Panoramic with 10 non-programmers. The results show that end-users can both understand and verify the behavior of complex location event specifications using Panoramic.

1 Introduction

Intelligent behavior in location-aware computing is driven by *location events*. Applications detect events by dynamically evaluating spatio-temporal relationships among people, places, and things. For example, a location-aware to-do list might detect simple events like “Alice is near the library” to trigger reminders. In contrast, many new applications for real-time location systems (RTLS) rely on *complex events* that contain sequences of interactions [1][27]. For example, a hospital workflow tracker may log a “cardiology exam” whenever a patient is detected exiting the hospital after meeting with a cardiologist and then spending time with a nurse and an electrocardiogram machine. With the RTLS market expected to soon exceed \$2 billion US [15], support for complex events is crucial.

A fundamental part of support for complex events is *event specification*; users must be able to specify new events to meet their evolving needs. Applications achieve this by leveraging event detection systems (e.g., context-aware computing infrastructures) that allow new events to be formally specified in some manner. However, while existing systems allow developers to specify new events using low-level APIs [3,30,33] or a declarative query language [10,14], dependence on developers is a costly inconvenience for both individuals and organizations. Indeed, a recent survey of location systems in hospitals cited the cost of tuning vendor software to site requirements as a critical barrier to deployment [34]. A compelling alternative is to allow direct specification of complex events by end-users. This is challenging, however, because it requires translation of high-level concepts into conditions on diverse, low-level, and uncertain sensor data. Unfortunately, existing systems for end-user event specification are either limited by inexpressive interfaces [17,24] or require iterative and potentially unfeasible training demonstrations for machine learning models [8].

It is equally important that end-users be able to *verify* event specifications and debug those that do not work. Verification is difficult because it requires a system to produce high-level evidence of a specification’s behavior over sensor traces that may be too complex for an end-user tool to generate. Moreover, because bugs can occur at the sensor level (e.g., calibration errors) or in the specification design, users must be able to understand detected events and evaluate their relationship to sensor data. This is impractical or impossible when events are specified with inscrutable machine learning models or when they do not represent uncertainty. As such, existing systems for end-user event specification are limited by inadequate support for verification and debugging.

We present Panoramic, a web-based tool that enables end-users to specify and verify complex location events. Panoramic does not require users to write code, understand complex models, perform elaborate demonstrations, generate test traces, or blindly trust deterministic events. Instead, it offers an intuitive visual language for specification and an intelligible verification interface that uses readily available historical sensor data. Specifically, we contribute:

1. A significant upgrade to our existing event detection system, Cascadia [37]; we facilitate event specification and verification by integrating Lahar [28], a new type of probabilistic event detector (Section 3).
2. Extensions that prevent errors and increase the expressive power of Scenic, our existing event specification tool. Our changes are guided by a formative user study with 12 non-programmers (Sections 3 and 4).
3. A novel approach to verification that leverages both historical sensor traces and a user’s knowledge of past events while explicitly but intuitively representing the probabilistic nature of sensor data and events (Section 5).
4. Verification interface widgets that provide end-users with an intelligible and hierarchical visualization of context. The widgets also allow users to distinguish sensor errors from errors in a specification’s design (Section 5).
5. A qualitative evaluation of event verification in Panoramic with 10 non-programmers. The study uses actual radio frequency identification (RFID) location traces collected from a building-scale deployment (Section 7).

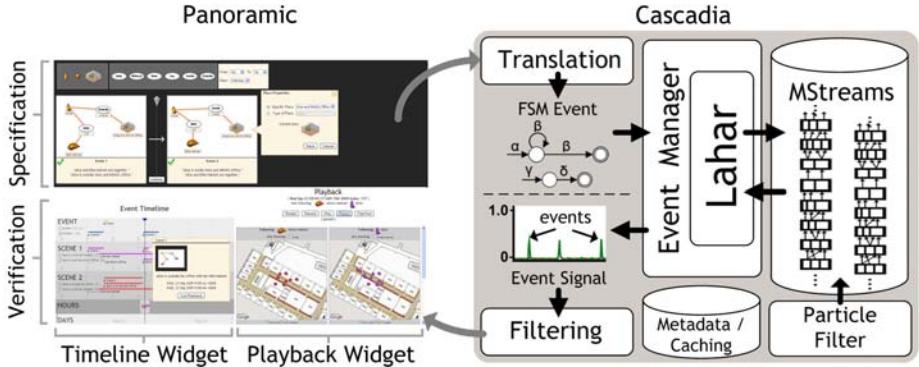


Fig. 1. The Panoramic system architecture. Events are iteratively specified in Panoramic, detected over historical location traces by Cascadia, and then displayed in Panoramic’s verification interface.

2 The RFID Ecosystem

We design and evaluate Panoramic using RFID traces, an increasingly common type of location data [29,34]. Our deployment, the RFID Ecosystem, models an enterprise RTLS deployment using 47 EPC Gen-2 RFID readers (160 antennas) installed throughout our 8,000 m² building. In addition, more than 300 passive (e.g., unpowered) tags carrying unique identifiers are attached to people and objects. When a tag passes an antenna it may wirelessly transmit its identifier to a reader, which in turn creates and sends a timestamped *tag read event* of the form (*antenna location*, *tag ID*, *time*) to a server for storage and processing. However, our antennas are only deployed in corridors (not inside offices), and past work has shown that factors like RF-absorbency and mobility in an everyday environment may prevent tags from being read [36,39]. Thus, like many location systems, the RFID Ecosystem may produce sporadic, imprecise location streams.

3 Specifying and Detecting Events

In this section, we describe Cascadia, Scenic, and extensions we make to support Panoramic. We also present a taxonomy for the events Panoramic can specify.

3.1 Integrating Lahar into Cascadia

Cascadia is a system for specifying, detecting, and managing RFID events [37]. It accepts declarative event specifications and detects the specified events over incoming RFID data, producing one event stream per specification. Cascadia copes with uncertainty by transforming intermittent RFID streams into smoothed, probabilistic *Markovian Streams* (MStreams) [19] that capture both the uncertainty of a tag’s location at each time step (e.g., a distribution over which rooms

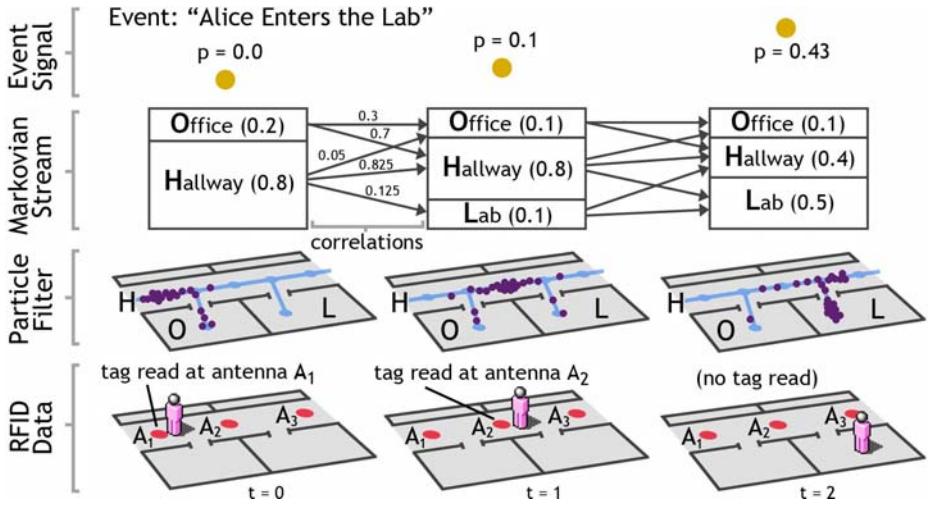


Fig. 2. Cascadia transforms raw, uncertain location data into smoothed, probabilistic Markovian streams over which Lahar detects complex events

the tag might be in) and the correlations between a tag’s possible locations (e.g., distributions over entire paths through a building). MStreams abstract away the complexities of sporadic and imprecise data to expose a more uniform model of location over which event specifications are expressed, thus considerably simplifying the requirements of an event specification language. At the heart of Cascadia is the PEEX event detection engine, which uses an SQL-like event specification language and extracts probabilistic events from MStreams.

We upgrade Cascadia by replacing PEEX with the Lahar event detection engine. Lahar’s query language is based on regular expressions that are represented internally with finite state machines (FSMs). The language’s pattern matching constructs, together with standard query predicates, offer a more intuitive way to express sequential spatio-temporal events. This streamlines translation from Panoramic specifications into Lahar queries and provides end-users with an easily comprehended mental model of the event detection process. Lahar produces a single probabilistic query signal for each MStream it processes. The query signal for an event specification, or *event signal*, consists of timestep-probability pairs $\langle t, p \rangle$ which indicate that the event occurred at timestep t with probability p (see Figure 2). Each probability p is derived from an MStream using well-established query answering techniques for probabilistic databases [19][28].

Lahar is an evolving research prototype that is currently limited to answering event queries over a single MStream. Panoramic, however, creates event specifications that reference multiple MStreams (e.g., multiple people and objects). As such, we developed a new *Event Manager* module for Cascadia that answers queries over multiple MStreams. The Event Manager translates specifications into sets of single MStream queries after which it orchestrates their execution

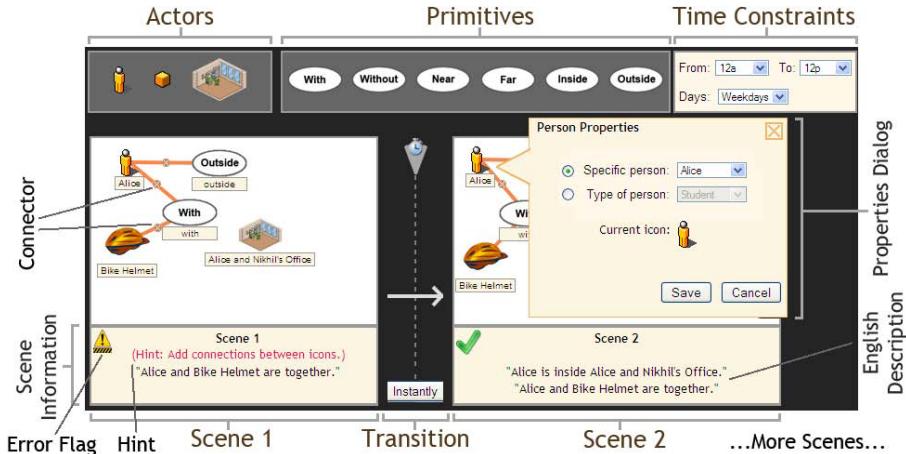


Fig. 3. The event specification interface employs a storyboard metaphor

with Lahar and merges the resulting event signals. This module also manages metadata on MStreams and caches intermediate event signals for reuse when answering other queries. Overall, our upgrades make Cascadia more expressive and provide cleaner semantics because all queries map to state machines.

3.2 Enhancing Scenic

Scenic is a tool that allows end-users to specify events for PEEX [37,38]. It uses an iconic visual language designed to support common location events and their composition through sequencing and conjunction. A storyboard layout describes how people and objects enact an event through a sequence of movements between places (see Figure 3). Users drag and drop *Actors* (people, objects and places) and *Primitives* (instantaneous events) onto *Scenes*. A *Sequence* of Scenes specifies a complex event as a sequence of spatio-temporal *sub-events*. Actors are specified using other end-user tools discussed in prior work [38].

While end-users understood Scenic [37], translation of specifications into PEEX queries was complex and imposed awkward constraints on the interface. For example, only certain combinations of Primitives could appear in a Scene and transitions between Scenes had to be of fixed length (e.g., one second). By designing Panoramic to target Lahar, we were able to remove these constraints and add several new features that increase expressiveness (see Figure 3). We added explicit *Connectors* between Actors and Primitives, enabling any combination of Primitives to appear within a single Scene. We also added *Transitions* between Scenes that are set explicitly by the user as occurring either *instantly* (i.e., in one timestep) or *over time* (i.e., some time may pass before the next scene occurs). Finally, we included simple constraints on absolute time (e.g., “before 12pm on Weekdays”) as a convenience.



Fig. 4. A Single Scene event: “*I’m in my office*”, translates into a single state FSM that enters the accept state whenever the user is inside his office

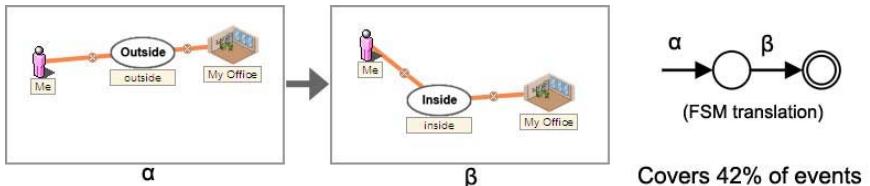


Fig. 5. A Consecutive Sequence event: “*I enter my office*”, translates into a linear FSM that enters accept state when Scenes are satisfied consecutively

3.3 Supported Events

In a survey of location events in pervasive computing applications [37] we found that six instantaneous events were used most often: *with*, *without*, *inside*, *outside*, *near*, and *far*; and that these events were frequently composed using *conjunction*, *repetition*, and *sequencing*. A large fraction of these events can be expressed in Panoramic and translated directly into FSM queries for Lahar. We now describe and illustrate the set of location events Panoramic supports.

Single Scene Events. Single Scene events use connected Primitives and Actors to describe a set of instantaneous events that occur simultaneously. Scenes including one person or object (i.e., one MStream) can be translated into a FSM query having a single accept state and a single incoming edge that is satisfied when the Scene’s Primitives are true (see Figure 4). When the Scene includes multiple people or objects, the Event Manager breaks the query into multiple single-state FSM queries, answers each, and merges the results by assuming their mutual independence and computing their conjunction. The set of events that can be represented by a single Scene is greatly extended by Connectors. Single Scene events form the basis of all location-aware computing applications and are sufficient to account for 24% of events recorded in the survey.

Consecutive Sequences. Consecutive Sequences contain Scenes separated by instantaneous Transitions. They are translated into linear FSM queries that have a single accept state preceded by a sequence of states and edges as shown in Figure 5. In the case of multiple objects or people, the Event Manager processes each Scene as though it were a Single Scene event, joins the results into a composite stream of independent events, and then runs a linear FSM query that corresponds to the Sequence. Consecutive Sequences are useful for detecting

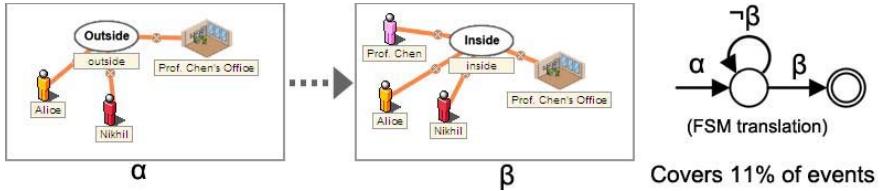


Fig. 6. A Sequence event with a gap: “Alice, Nikhil and Prof. Chen begin a meeting in Prof. Chen’s office”. This event occurs as soon as the last participant enters the office. A FSM with a self-loop is used to wait for the last participant.

state transitions like entering or exiting a place, approaching an object, or beginning a trajectory; such events account for another 42% of surveyed events.

Sequences with Gaps. Sequences may also contain Transitions that allow time (and potentially other events) to pass between one Scene and the next. Panoramic translates these Sequences into linear FSMs having a self-loop edge that is satisfied by the *negation* of the condition on the edge which leads to the next state. For example, the self-loop edge in Figure 6 ensures that the FSM remains in “pre-meeting” state until Alice, Nikhil, and Prof. Chen are all in Prof. Chen’s office. Sequences with gaps represent a class of events not previously supported by Cascadia, thus increasing the flexibility and expressiveness of specifications. They also account for another 11% of the events in our survey.

Disjunctions of Sequences. A disjunction of Sequences may represent different orderings of instantaneous events or alternate paths in a workflow. While such events comprise less than 5% of events in the surveyed literature, they are of growing importance in emerging domains like workflow monitoring in hospitals [27]. Panoramic does not directly support disjunctions, but users may specify multiple Sequences and compose their disjunction within an application. Here we stress that Cascadia has the capability to support disjunctions of Sequences and that we plan to extend Panoramic to directly specify them in future work.

Actor variables also create disjunctions of Sequences. Most surveyed events can be usefully modified with variables. For example, instead of “Alice meets Nikhil in her office”, one could express “Two researchers meet in any room”. These events are translated into a disjunction of Sequences where the variables in each Sequence are replaced with a different combination of possible values. This is effective for events with few variables that range over a limited domain.

Unsupported Events. There are another 18% of surveyed events for which Panoramic provides limited or no support. This includes repeating sequences, which can be translated into FSMs with cycles. These queries are computationally difficult to answer over probabilistic data, but can be approximated by unrolling the cycle. Other common and unsupported location events involve precise 3D location and events involving speed of transition (i.e., velocity).

3.4 Formative Evaluation

In addition to the move from PEEX to Lahar, the design of Panoramic’s event specification interface was driven by a formative user study. In this study, 12 non-programmers were given a brief tutorial on Panoramic and asked to complete a series of event specification tasks while talking aloud. Each task presented participants with an example application and usage scenario for which they were asked to specify an appropriate event. After participants declared a task complete, a researcher would then review the produced specification and explain exactly how Panoramic would interpret it. Participants could then revise their specification if the researcher’s explanation did not match their intention.

As a result of the study, we identified a variety of usability and expressivity problems which we addressed with the design features presented above. However, we also observed several more fundamental problems regarding user ability to understand and verify the behavior of a created specification. We discuss these problems in the next section.

4 Problems in Specifying an Event

Even with the enhancements to the Scenic interface, users encountered a variety of difficulties while authoring an event specification with Panoramic. Here, we summarize and discuss those most frequently observed in our formative study.

4.1 Syntax Errors

Nearly all participants produced one or more syntactically illegal specification. Most syntax errors were the result of forgotten connections or targeting errors while rapidly creating a Scene. In a few cases, participants made mistakes because they did not clearly understand how Actors and Primitives could be connected. We addressed these problems with three new features (see Figure 3). First, we constrained the interface to allow only legal connections between Actors and Primitives, thereby eliminating the possibility of syntax errors. We also added an *information panel* below each Scene that displays English explanations for fully connected Primitives in that Scene. Finally, we included a *status flag* in the information panel that shows a green checkmark when a Scene is legally specified or an under construction symbol when more work needs to be done. If more work is needed, the information panel provides a hint as to what elements (e.g., Actors, Primitives, Connectors) are needed to complete the Scene.

4.2 Design Problems

A problem encountered by 3 of the first 5 participants was in deciding on a specification design that would meet the task’s requirements. While 1 participant had difficulty reasoning about what needed to be specified, others knew what they wanted to specify but weren’t sure how the available widgets could be composed to do it. To address both of these problems, we introduced *event*

templates, stock specifications that provide examples of common events with their usage scenarios - at least one for every type of event in the taxonomy. The last 7 participants in our formative study were provided with a library of templates during the study session; those that encountered design challenges were able to use the template library to decide on a design.

4.3 Problems with Timing

Half of the participants chose to revise a completed specification due to problems with timing. These problems consisted in use of the wrong Transition type (e.g., *instantly* instead of *over time*) or in using one Scene when two Scenes were needed. For example, a participant intended to specify “the custodian leaves the lab and goes to the closet” as two Scenes separated by an instantaneous Transition. This specification was flawed because a custodian cannot move instantly between rooms. Another participant used a single Scene to tell a context-aware notifier to send her an email when a meeting occurs. While the single Scene would effectively detect a meeting and send an email, it would also occur repeatedly throughout the meeting, causing many emails to be sent. Event templates helped to reduce problems with timing, but some more subtle problems remained, forcing participants to revise their specifications. We therefore developed additional solutions to timing problems which we present in Section 5.

4.4 Tuning the Level of Specificity

A critical challenge faced by all participants while designing and revising their specifications was to determine the appropriate level of specificity. Some participants routinely *under-specified* events by leaving out Actors or Primitives. One explained her under-specification by saying “I wanted it to cover every case so I only need one event,” another simply explained that it took two revisions before he realized that another object was needed. *Over-specification* was also common. In such cases, participants often explained that they added non-essential Actors, Primitives, or Scenes because they weren’t confident that Panoramic would detect the intended event without them. For example, one participant constrained a “group meeting” event to occur only when all group members were together in a room with laptops and coffee mugs. He explained that “they could be in the room for some other reason, but if they all have laptops and coffee then they’re probably in a meeting.” Whether an event is over or under specified, the result is a specification that does not fit the user’s intention.

While mismatches in specificity may be less of a problem when users are reasoning about events in their own life with which they are intimately familiar, some tuning is likely to be required whenever a new event is specified. Many participants adopted a trial and error methodology when specifying an event, using the researcher’s explanation to “test” an event’s behavior over multiple design iterations. This motivated the design of additional interface components that support users in understanding and verifying the behavior of a specification. We discuss these components in the next section.

5 Understanding and Verifying Events

In response to the problems discussed in the previous section, we extended Panoramic to support end-users in understanding and verifying their event specifications. Here we face the hard problem of generating test data for specifications. Our approach is to allow users to assess the correctness of a specification by running it on *historical data*. We use Cascadia to detect the event together with the timeline-based and map-based widgets presented below to visualize the results. The advantage of this approach is that it does not require complex simulations or synthetic data which may not be truly representative. Instead, it reveals the behavior of the event on real, readily available sensor data. The obvious constraint is that the tested event must have already been recorded by user’s RFID deployment. This is a reasonable sacrifice for scenarios like hospitals and office environments where both simple events and complex workflows occur repeatedly.

5.1 Timeline Overview

We developed a timeline widget (see Figure 7) that provides a rapid overview of detected event results as horizontal *bars* in a timeline where a bar’s start and end points correspond to a detected event. Events are organized in groups of sub-events (e.g., sub-sequences, Scenes, Primitives) in order to provide an in-depth explanation as to why an event was or was not detected. Each group of events is displayed in its own *band*, with one band for the event itself along with its sub-sequences, and one band for each Scene with its Primitives. By correlating the sub-event bars with the presence or absence of an event bar, users can gain an understanding of how each sub-event contributes to or detracts from the detection of the event. For under-specified events, this process can reveal that frequently occurring sub-events result in a larger than intended number of detections. In the case of over-specification, it can show a user that absent sub-events are preventing the event from being detected. Timing problems are also visible as unexpectedly long or short event durations.

The timeline further facilitates the verification process with exploratory browsing functions. The timeline can be dragged left or right to move through time, and two zoomed-out bands (one for hours and one for days) provide additional context for large datasets. By default, event bands are rendered with minute-level granularity but may be zoomed in or out using the mouse scroll wheel. Checkbox labels on the left side of each band describe the contents of each sub-event group in that band. The bars representing a sub-event can be shown or hidden from the timeline view by checking or unchecking that sub-event’s checkbox. Selecting a checkbox label will highlight all occurrences of an event and its sub-events in the timeline. Finally, clicking a bar in the timeline brings up a bubble containing a thumbnail image of the corresponding event or sub-event along with an English description and precise timing information.

The timeline-based design is particularly well-suited for display of sequential data and allows users to leverage their natural ability for reasoning about temporal events [16]. With a suitably large dataset, we anticipate that the timeline view can help users to quickly gather evidence of a specification’s behavior.

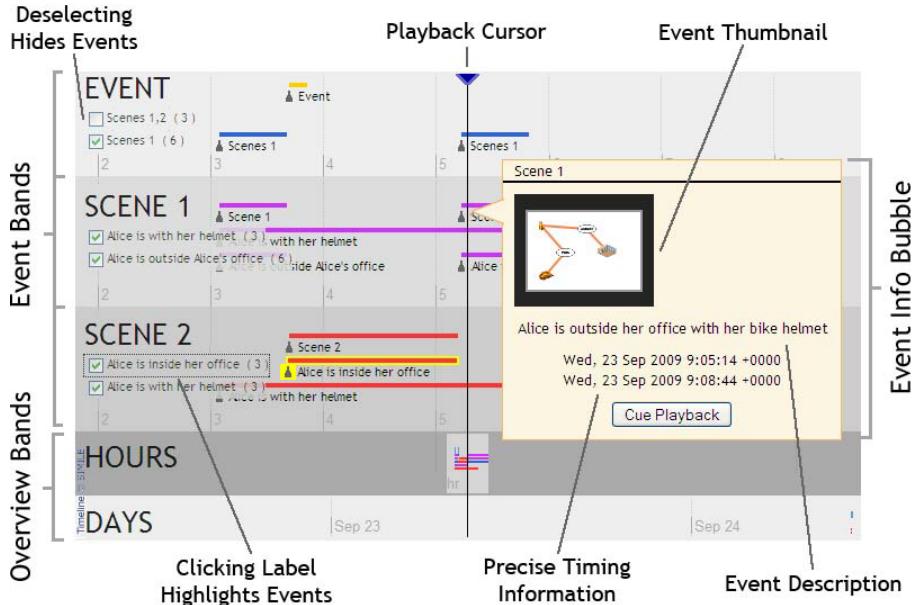


Fig. 7. The timeline reviews the event “Alice enters her office with her helmet”

Filtering Event Signals. The timeline must present a simple, discrete view of the complex probabilistic data it displays. As such, we take several steps to transform raw event signals from Lahar into discrete event streams that are amenable to visualization. First, because the probability of a true event occurrence may vary widely, we identify and flag all local maxima in a signal as potential event occurrences. We then filter out all spikes (i.e., peaks lasting less than 2 seconds) from this set. Spikes are unlikely to correspond to a true event because they are faster than any action humans commonly perform. They may occur in an entered-room event signal, for example, when a user passes but does not enter the room. After removing spikes, we transform the remaining “humps” into discrete events having the same duration and which can be displayed on the timeline.

5.2 Detailed Playback

Though it provides a useful overview and a direct look at the cause for Sequence and Scene events, the timeline does not explain why a given *Primitive event* does or does not occur. This is a crucial question when working with real historical sensor traces because missed RFID readings can lead to false positives or negatives that are indistinguishable from unexpected behavior in the timeline. For example, a “group meeting” event may match a user’s intention but fail to work in practice because one group member’s RFID tag is routinely missed. Using only the timeline, it would be impossible to distinguish an absent group member from a tag that needs to be replaced. To help users identify problems that are rooted in sensor errors rather than in a specification, we developed a map-based

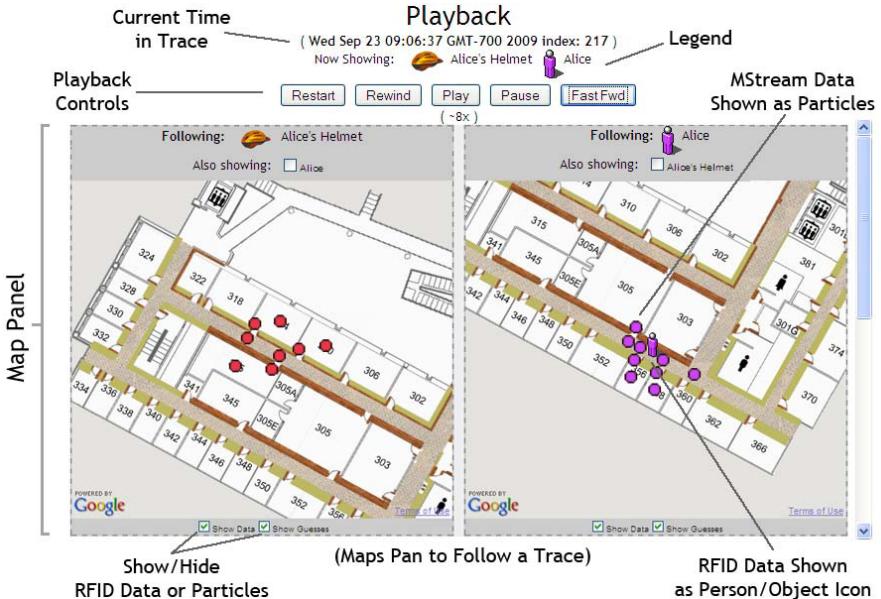


Fig. 8. The playback widget reviews traces for Alice and her helmet. Person and helmet icons represent raw location data. Particle icons display a probability distribution over a tracked entity’s possible locations.

trace playback widget. The widget allows a user to semantically drill-down into any point in the timeline and review both the sensor trace and the MStream starting at that time.

The playback widget renders at the right side of the timeline (see Figures 1 and 8) whenever a user clicks the “Cue Playback” button in the pop-up bubble for a timeline bar. At the same time, a playback cursor appears over the timeline to designate the current time in the trace to be replayed. Users may also be interested in portions of the timeline that contain no detected events, as such, they can drag the timeline to any location to cue playback there. Standard video controls (e.g., pause, play, stop, rewind and fast-forward) provide a familiar interface for reviewing segments of the trace. Playback occurs in a collection of *map panels* that show RFID readings and MStream data (i.e., particles from the particle filter) overlaid on a map of the RFID deployment. The timeline scrolls synchronously with the map-based playback. Each such panel follows a particular person or object from the trace, automatically panning and switching floors as needed. The user may also choose to show multiple traces in a single map panel by selecting checkboxes above the map that correspond to additional traces. RFID readings and MStream data may be switched on and off in a given map panel using the “Show Data” and “Show Model” checkboxes. Unneeded map panels can be collapsed to facilitate side-by-side comparison of traces. Labels at the top of the playback widget show the current time in the trace and summarize the collection of people and objects being viewed in the trace.

6 Implementation

Panoramic’s event specification and verification interfaces are entirely web-based and built using the Google Web Toolkit [12]. The playback widget was built using Google Maps [11], and a customized version of MIT’s Simile Timeline [31] was used to implement the timeline widget. In total, Panoramic contains 152 Java classes that control the interface and orchestrate AJAX communication with a server. An additional 42 classes were added to Cascadia to support translation and execution of Panoramic events with Lahar.

7 Evaluation

We ran a qualitative study of Panoramic’s event verification capabilities with 10 non-programmers. Participants were offered \$20 to perform 60 minutes of event verification tasks. In addition to a 10 minute tutorial, participants were prepared with a background story that described a week in the life of Alice, a fictional student. The story included precise information on events that occurred (e.g., “group meeting on Monday at 11”) as well as information on events that often occur (e.g., “Nikhil often stops by Alice’s office to talk”). Each task included a specification, a description of the application and usage scenario for which Alice created the specification, and the corresponding detected events from the week of Alice’s data. Participants were asked to talk aloud as they used Panoramic in combination with what they knew about Alice’s week to decide how each specification worked, whether it met her needs, and how it could be fixed.

The week of historical data was spliced together from traces collected in the RFID Ecosystem and for which we have ground truth information on when and where events occurred. We combined a set of high fidelity traces with a small number of highly ambiguous traces (e.g., traces with a large number of missed tag reads). The first four tasks were presented in random order with one task having a specification that clearly succeeded over the week of data and three that failed because they were over-specified, under-specified, or contained a timing error. A fifth task contained a specification that should have met Alice’s needs but was not detected over the week’s data as a result of ambiguous sensor traces.

7.1 Observations and Enhancements

Overall, participants were able to complete the tasks, averaging 15-20 minutes for the task involving ambiguous traces and 10-15 minutes for all other tasks. All participants understood the behavior of specifications and could distinguish sensor errors from specification errors. Participants were also able to grasp the intended behavior of a specification both from the usage scenario and from the specification itself. As such, they used the timeline and playback widgets as a means for verifying intuitions about a specification’s behavior rather than for exploring its overall behavior. Moreover, though overconfident participants initially declared a flawed specification to be correct in 6 of 50 tasks, they quickly changed

their minds after comparing the timeline to their knowledge and intuitions about events. All participants were comfortable using Panoramic and several remarked that it was fun or “like a game”. However, while they did not encounter any critical barriers to task completion, many participants faced recurring difficulties for which we have developed preliminary solutions (see Figure 9).

First, participants often checked the timeline for consistency with the events they knew occurred during Alice’s week. In many cases the first question they tried to answer was “how many times was the event detected?”. The timeline does not directly answer this question, so participants had to scroll through the week to count event occurrences. We addressed this problem by adding a count for each event beside that event’s label at the left side of the timeline.

The task involving an under-specified event required participants to further constrain the specification by adding an object. This was difficult because the set of available objects was buried in the specification interface, leaving participants unsure of what objects were available. Moreover, without the ability to review traces for other Actors alongside the currently loaded trace, it was difficult to decide whether or not another Actor was relevant to an event. While this problem may be less critical when users reason about people and objects they know, we did introduce a new section to the legend at the top of the playback panel that shows other Actors which may be relevant to the event. By clicking the checkbox next to an Actor, users can load a new map panel that plays the trace for that Actor. The set of displayed Actors is currently chosen as those that are proximate to, or move in the same time window as the currently loaded trace. This is a reasonable compromise to the unscalable alternative of displaying every possible Actor because proximate or moving Actors are likely to be more relevant.

Two additional recurring frustrations were voiced by participants when using the playback widget. First, they had difficulty understanding the visualization of the MStream as a set of particles. After explaining the particles as “Panoramic’s guess at where a person or object is,” participants were better able to understand but still had difficulty reasoning about sensor errors and missed detections as a result of ambiguity. As such, we changed our rendering of particles to include an

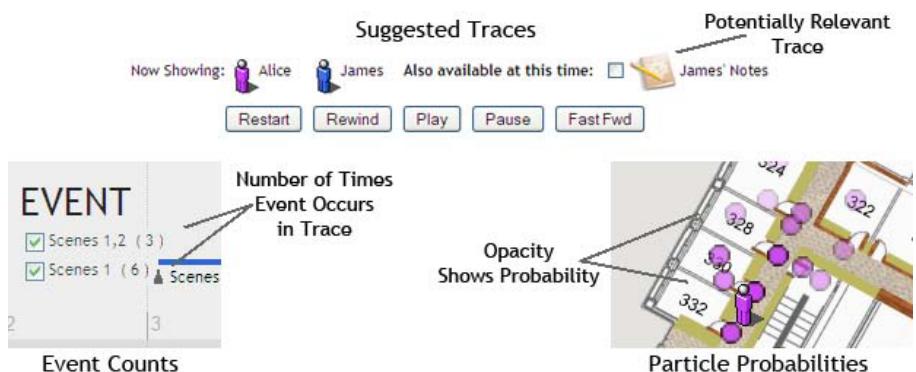


Fig. 9. Enhancements to the timeline and playback widgets

opacity level that corresponds to a particle’s probability. This helped the last 8 of 10 participants to identify sensor errors in the ambiguous trace 3-5 minutes faster than the 2 who did not have this feature available. A second difficulty was that participants felt it was difficult to correlate the trace playback with the events in the timeline. Although the timeline was animated to correspond to the trace, participants were uncomfortable looking back and forth between the trace and the timeline. One participant explained her difficulty with the playback widget by saying “it shows where Panoramic thinks the people are, but not what it thinks about the events, you have to keep looking back at the timeline to see the events, that’s hard”. This problem could be addressed in future work by arranging the timeline below the playback widget, or by embedding pop-ups in the playback maps that mark when and where events occur.

7.2 Limitations

Panoramic currently has several key limitations. First, while it supports specifications with Actor variables, verification of such events is tedious because users may need to review multiple sets of historical traces - one for each possible parameterization of the variables. Panoramic is also limited by its minimal support for debugging suggestions. While a list of potentially relevant Actors is useful, more intelligent suggestions could be made by assessing the how sensitive the event detection results are to slight changes in a specification. Finally, Panoramic’s reliance on historical data may be problematic for events that seldom occur. Here it may be possible to automatically generate synthetic test data for a particular specification using techniques similar in spirit to recent work by Olston *et al.* [25]. Future work will address these and other limitations.

8 Related Work

Here we review and discuss a variety of end-user software engineering techniques for sensor systems. We focus on specification and verification of events, omitting discussion of techniques that specify event-triggered behaviors.

Specification Languages. Event specification systems for end-users are difficult to design because they must lower the barrier to entry without compromising expressive power. One approach is to create a specification language with abstractions that represent high-level concepts in the target application domain. For example, early systems like PARCTAB [35], Stick-e Notes [26], and SPECs [17] used scripts to describe primitive location events. More recent work with Semantic Streams [40] and probabilistic context-free grammars [22] can detect some complex and even uncertain events in sensor networks. While these systems use data processing techniques similar to those in Cascadia, they expose languages that are not well-suited to end-users.

Specification Interfaces. Several systems have made specification more accessible with graphical interfaces that declaratively specify events. EventManager [24] used four drop-down boxes to specify a small set of primitive location events. CAMP [33] specifies non-sequence (i.e., instantaneous) events with a

magnetic poetry interface that answers the questions: who, what, where and when. The Topiary [20] design tool also specifies instantaneous events, but uses an interface with an active map and a storyboard. Panoramic is quite similar to iCAP [32], a visual interface for specifying spatio-temporal sequence events. However, CAMP, iCAP, and Topiary are all less expressive than Panoramic because they rely on custom-coded event detection modules instead of a flexible event detection engine like Lahar. Moreover, they do not explicitly support event detection over uncertain sensor data like Panoramic.

Programming by Demonstration. Another approach is programming by demonstration (PBD), in which users supply example sensor traces to train an event detector. a CAPpella [8] is a PBD system that allows users to train a Dynamic Bayesian Network with labeled sensor traces. Apart from low detection rates, a CAPpella is difficult to use in our motivating scenarios because it requires 5 or more traces for training. Other systems focus on detecting simpler events with fewer examples and rapid feedback. For example, Crayons [9] and Eye-patch [23] enabled users to rapidly train visual classifiers using a demonstration-feedback loop. The Exemplar system [13] employs a similar loop featuring an algorithm that requires only one demonstration. Exemplar focuses on exposing an intelligible and editable visualization of its model, addressing the fact that automatically learned models are often inscrutable [5][18]. These prior PBD methods become impractical for complex events that involve the simultaneous movement of people and objects.

Verifying Specifications. Past work has shown that end-users must be able to verify that a specification works as intended [5][18]. Verification identifies three broad categories of error: (1) syntactic errors that make specifications illegal or ambiguous, (2) semantic errors that make valid specifications behave in unexpected ways, and (3) sensor errors that cause event detectors to erroneously detect or miss events. Most languages and declarative interfaces use interactive visual feedback (e.g., error flags, prompts for disambiguation) to cope with syntactic errors [20][32]. Semantic errors are often identified by testing with sensor traces (as discussed below). However, both CAMP and Panoramic can reveal semantic errors in non-sequence events by generating high-level English descriptions - in Panoramic these are descriptions of Primitives. Most systems provide no support for identifying sensor problems beyond what may be inferred from detection errors. In contrast, Panoramic directly supports discovery of sensor errors by providing a visualization that correlates sensor data with detected (and missed) events.

Trace-Driven Debugging. Test traces are commonly generated using either Wizard of Oz [20][32], in which the user simulates sensor traces with a special interface, or demonstration [8][13], in which the user enacts an event while recording it with sensors. Both of these techniques become prohibitively demanding for complex events. The Wizard of Oz approach also fails to capture the impact of uncertainty in real sensor data. Panoramic avoids these problems by using pre-recorded traces. Moreover, though other systems could adopt Panoramic's approach, they do not provide the support for archiving, exploration, and visualization of uncertain

events and sensor data that Panoramic does. Debugging also requires that users correct erroneous specifications. This is a simple matter of modifying the specification in declarative interfaces like iCAP and Panoramic. It is much less straightforward in PBD systems [5], and may require that entirely new sets of demonstrations be recorded.

Intelligible Context Models. Several papers have established and articulated the need for intelligible models of context [16,21]. A few systems have also explored support for intelligible context. Cheverst *et al.* [6] supported users with scrutable decision tree rules and context histories. The PersonisAD [2] framework allowed developers to access supporting evidence for context items. Dey and Newberger [7] support intelligibility in the Context Toolkit using Situation components that expose application logic to developers and designers. Panoramic adds to this body of work by providing an intelligible, scrutable context model for complex location events. Moreover, Panoramic contributes a context management system that directly copes with uncertainty using probabilities while not requiring users to explicitly specify probability thresholds.

9 Conclusion

In this paper we presented the design and evaluation of Panoramic, an end-user tool for specifying and verifying RFID events. Our design leverages and extends the Cascadia system and the Scenic tool in significant ways, and is informed by feedback from a formative study with 12 non-programmers. We also contributed the design of an interface for verifying complex location events that was motivated by problems observed in the formative study. We evaluated our verification interface with 10 non-programmer study participants and found that in spite of minor difficulties with the interface, all users were able to complete five representative verification tasks. Overall, we have presented a tool that satisfies a growing need in a way that is accessible to end-users and which works in spite of inevitable sensor errors. Moreover, we demonstrated techniques that support intelligible context for applications that use complex location events.

Acknowledgements

The authors would like to thank Kayla Gould and Emad Soroush for their efforts in support of the design and evaluation of Panoramic. This work was supported in part by the National Science Foundation under grants CNS-0454394, CNS-0454425, IIS-0713123, and IIS-0812590; and by gifts from Intel Research and the University of Washington's College of Engineering.

References

1. Amelior ORTracker: Orchestrate Patient Flow (2009),
<http://www.pcts.com/unified/ortracker.php>
2. Assad, M., et al.: PersonisAD: Distributed, Active, Scrutable Model Framework for Context-Aware Services. In: LaMarca, A., Langheinrich, M., Truong, K.N. (eds.) Pervasive 2007. LNCS, vol. 4480, pp. 55–72. Springer, Heidelberg (2007)

3. Bardram, J.E.: The Java Context Awareness Framework (JCAF) – A service infrastructure and programming framework for context-aware applications. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) PERVASIVE 2005. LNCS, vol. 3468, pp. 98–115. Springer, Heidelberg (2005)
4. Bellotti, V., Edwards, K.: Intelligibility and Accountability: Human Considerations in Context Aware Systems. HCI 16, 193–212 (2001)
5. Chen, J., Weld, D.S.: Recovering from Errors During Programming by Demonstration. In: IUI 2008, pp. 159–168 (2008)
6. Cheverst, K., et al.: Exploring Issues of User Model Transparency and Proactive Behaviour in an Office Environment Control System. User Modeling and User-Adapted Interaction 15(3-4), 235–273 (2005)
7. Dey, A., Newberger, A.: Support for Context-Aware Intelligibility and Control. In: CHI 2009, pp. 859–868 (2009)
8. Dey, A.K., et al.: A CAPpella: Programming by Demonstration of Context-Aware Applications. In: CHI 2004, vol. 1, pp. 33–40 (2004)
9. Fails, J., Olsen, D.: A Design Tool for Camera-Based Interaction. In: CHI 2003, pp. 449–456 (2003)
10. Garofalakis, M.N., et al.: Probabilistic Data Management for Pervasive Computing: The Data Furnace Project. IEEE Data Eng. Bull. 29(1), 57–63 (2006)
11. Google Maps API - Google Code, <http://code.google.com/apis/maps/>
12. Google Web Toolkit - Google Code (2009), <http://code.google.com/webtoolkit/>
13. Hartmann, B., et al.: Authoring Sensor-Based Interactions by Demonstration with Direct Manipulation and Pattern Recognition. In: CHI 2007, pp. 145–154 (2007)
14. Heer, J., et al.: Liquid: Context-Aware Distributed Queries. In: Dey, A.K., Schmidt, A., McCarthy, J.F. (eds.) UbiComp 2003. LNCS, vol. 2864, pp. 140–148. Springer, Heidelberg (2003)
15. Real-time locating systems 2009-2019 (2009),
<http://www.idtechex.com/research/reports/>
16. Knoll, S., et al.: Viewing Personal Data Over Time. In: CHI 2009 Workshop on Interacting with Temporal Data (April 2009)
17. Lamming, M., Bohm, D.: SPECs: Another Approach to Human Context and Activity Sensing Research, Using Tiny Peer-to-Peer Wireless Computers. In: Dey, A.K., Schmidt, A., McCarthy, J.F. (eds.) UbiComp 2003. LNCS, vol. 2864, pp. 192–199. Springer, Heidelberg (2003)
18. Lau, T., et al.: Why PBD Systems Fail: Lessons Learned for Usable AI. In: CHI 2008 (2008)
19. Letchner, J., et al.: Challenges for Event Queries over Markovian Streams. IEEE Internet Computing 12(6), 30–36 (2008)
20. Li, Y., et al.: Topiary: A Tool for Prototyping Location-Enhanced Applications. In: UIST 2004, pp. 217–226 (2004)
21. Lim, B., Dey, A.: Assessing Demand for Intelligibility in Context-Aware Applications. In: Ubicomp 2009, pp. 195–204 (2009)
22. Lymberopoulos, D., et al.: A Sensory Grammar for Inferring Behaviors in Sensor Networks. In: IPSN 2006, pp. 251–259 (2006)
23. Maynes-Aminzade, D., et al.: Eyepatch: Prototyping Camera-Based Interaction Through Examples. In: UIST 2007, pp. 33–42 (2007)
24. McCarthy, J.F., Anagnost, T.D.: EVENTMANAGER: Support for the Peripheral Awareness of Events. In: Thomas, P., Gellersen, H.-W. (eds.) HUC 2000. LNCS, vol. 1927, pp. 227–235. Springer, Heidelberg (2000)
25. Olston, C., et al.: Generating Example Data for Dataflow Programs. In: SIGMOD 2009, pp. 245–256 (2009)

26. Pascoe, J.: The Stick-e Note Architecture: Extending the Interface Beyond the User. In: IUI 1997, pp. 261–264 (1997)
27. Philly Hospital Uses RTLS to Track Patient Flow, Care and Training (May 2009), <http://www.rfidjournal.com/article/view/4934/1>
28. Ré, C., et al.: Event Queries on Correlated Probabilistic Streams. In: SIGMOD 2008, June 2008, pp. 715–728 (2008)
29. RTLS Providers Cite Strong Demand From Hospitals (June 2009), <http://www.rfidjournal.com/article/print/4981>
30. Salber, D., et al.: The Context Toolkit: Aiding the Development of Context-Enabled Applications. In: CHI 1999, pp. 434–441 (1999)
31. SIMILE Timeline (2009), <https://simile.mit.edu/timeline/>
32. Sohn, T., Dey, A.: iCAP: An Informal Tool for Interactive Prototyping of Context-Aware Applications. In: CHI 2003, pp. 974–975 (2003)
33. Truong, K.N., et al.: CAMP: A Magnetic Poetry Interface for End-User Programming of Capture Applications for the Home. In: Davies, N., Mynatt, E.D., Siio, I. (eds.) UbiComp 2004. LNCS, vol. 3205, pp. 143–160. Springer, Heidelberg (2004)
34. Vilamovska, A., et al.: Study on the requirements and options for RFID application in healthcare. Technical report, RAND Corporation (July 2009)
35. Want, R., et al.: An Overview of the PARCTAB Ubiquitous Computing Experiment. IEEE Personal Communications 2(6), 28–33 (1995)
36. Welbourne, E., et al.: Challenges for Pervasive RFID-based Infrastructures. In: PERTEC 2007, March 2007, pp. 388–394 (2007)
37. Welbourne, E., et al.: Cascadia: A System for Specifying, Detecting, and Managing RFID Events. In: MobiSys 2008, June 2008, pp. 281–294 (2008)
38. Welbourne, E., et al.: Building the Internet of Things Using RFID: The RFID Ecosystem Experience. IEEE Internet Computing (May 2009)
39. Welbourne, E., et al.: Longitudinal Study of a Building-Scale RFID Ecosystem. In: MobiSys 2009 (June 2009)
40. Whitehouse, K., et al.: Automatic Programming with Semantic Streams. In: SenSys 2005, November 2005, pp. 290–291 (2005)

Tactile Wayfinder: Comparison of Tactile Waypoint Navigation with Commercial Pedestrian Navigation Systems

Martin Pielot and Susanne Boll

OFFIS Institute for Information Technology, Germany
martin.pielot@offis.de,
susanne.boll@uni-oldenburg.de
<http://www.offis.de>

Abstract. In this paper we report on a field study comparing a commercial pedestrian navigation system to a tactile navigation system called *Tactile Wayfinder*. Similar to previous approaches the *Tactile Wayfinder* uses a tactile torso display to present the directions of a route's waypoints to the user. It advances those approaches by conveying the location of the next two waypoints rather than the next one only, so the user already knows how the route continues when reaching a waypoint. Using a within-subjects design, fourteen participants navigated along two routes in a busy city centre with the *Tactile Wayfinder* and a commercial pedestrian navigation system. We measured the acquisition of spatial knowledge, the level of attention the participants had to devote to the navigation task, and the navigation performance. We found that the *Tactile Wayfinder* freed the participants' attention but could not keep up with the navigation system in terms of navigation performance. No significant difference was found in the acquisition of spatial knowledge. Instead, a good general sense of direction was highly correlated with good spatial knowledge acquisition and a good navigation performance.

Keywords: Tactile Displays, Pedestrian Navigation, Wayfinding, Tactons.

1 Introduction

With the success of the iPhone and similar platforms the navigation software known from our cars has become available in our pocket (see Fig. 1). Offering routing modes for pedestrians we can have these applications guide us to unknown places, turning them into personal navigation devices (PND). The interaction with those applications has not changed much: our position is displayed on an interactive map and the route to the destination is highlighted. Additionally, we are given route instructions by speech, text, and visual cues.

Pedestrians, however, use these applications in different contexts than car drivers. No cage of steel is protecting them from environmental interferences, such as sun, rain, or noise. While walking it is hard to read a visual display and



Fig. 1. Using a PND for (pedestrian) navigation in an urban environment

pay attention to the environment at the same time. In addition, if e.g. sunlight reflects from the screen it can become difficult to identify anything. Using speech and sound can help, but auditory information via speakers can be missed (due to noise) or be socially inappropriate (if the user does not want to stand out). Headphones can solve both problems but cut the user off from the only sense that allows sensing potential threats all around the user, such as a car approaching from behind. Thus, audio-visual displays can be unsuited in many situations that pedestrians typically face when navigating.

Utilizing the sense of touch for navigation as a solution has been proposed by several groups [24,28,27,10]. These groups used tactile displays to convey the direction of the next waypoint or the destination. It has been shown that such systems can decrease the cognitive workload [6], and support the interpretation of geographic maps [22,18]. However, it has yet to be shown that tactile waypoint navigation can outperform traditional PNDs.

In this paper we report from a field study comparing this tactile waypoint navigation concept to a commercial PND. While we confirmed that waypoint navigation with tactile displays can free the users' attention, the navigation performance was worse compared to the PND. Beyond the navigation system we found that the user's sense of direction was a major factor for the navigation performance. We argue that therefore good navigation systems should support the sense of direction.

2 Related Work

Presenting route instructions on mobile devices has received a lot of attention in the community already. Kray et al. [13] discussed means for providing route instructions on mobile devices, including 2D and 3D maps. In a pilot study they

found that females prefer 3D maps, while males prefer 2D maps, which could, however, be attributed to the fact that the male participants had a higher level of experience with 2D maps. Ishikawa et al. [11] investigated the wayfinding behaviour depending what method was used to learn the route. They compared PNDs, paper maps, and learning the route by self-experience, i.e. being guided along it. PNDs performed significantly worse in terms of spatial knowledge acquisition, navigation performance, and subjective difficulty rating.

While recent PNDs present distances to inform the user's about the location of waypoints (e.g. enter round-about in 200m), humans tend to use landmarks instead [17]. Photos of landmarks combined with route instructions outperforms paper maps and reduces the mental workload [7]. Still this approach requires interacting with a tiny display. Rukzio et al. [21] therefore proposed presenting route instructions on public displays instead of the mobile device. The public display is used to point into the direction the pedestrian has to proceed. In a follow-up study [20] they could show that such public navigation displays can reduce the mental workload and the frustration level.

In order to overcome the problems with auditory and visual display Tan and Pentland [23] proposed the use of tactile display for navigation. Bosman et al. [2] showed that providing turning directions by two vibrotactile actuators worn at the wrists outperformed following signposts in an indoor navigation task. Tsukada et al. [24] proposed a tactile torso display called ActiveBelt for - amongst other things - waypoint navigation. The display consists of an array of vibrotactile actuators attached to a belt. When it is worn around the waist, the actuators get equally distributed around the torso. The vibrotactile signals produced around the torso can intuitively be interpreted as horizontal directions [25]. Field studies [28][27] showed the feasibility of such tactile belts for waypoint navigation. It could also be shown that the tactile modality reduces the overall cognitive load and improves situation awareness compared to visual user interfaces [6][22].

It has yet to be shown that tactile displays can overcome the issues of audio-visual interfaces regarding pedestrian navigation systems in urban environments. Existing studies either lack a baseline [24][28][27], compare tactile waypoint navigation with other aids than route-instructions [2][22], or were conducted in non-urban environments [6].

3 Limitations of Today's Navigation Systems for Pedestrian Navigation

Navigation systems for pedestrians still employ the same visual interaction metaphor as car navigation systems. Figure 2 shows an example of a commercial navigation application with a pedestrian mode. The user's position is shown on a street map. The route to the destination is highlighted. The map is aligned so the forward direction of the user's movement corresponds to "up" in the display. Speech output is used to announce the distance to the waypoint and the proposed turning direction in regular intervals. To overcome the positioning technologies' inherent inaccuracy, the user's position is approximated by a technique called *map matching*. Assuming that cars typically will not leave the road

the technique maps the GPS position onto the nearest street, in cases where the satellite signal is inaccurate. While these systems are quite successful in cars, recent user studies have highlighted limitations that arise due to the situational context of pedestrians.

Spatial Knowledge Acquisition. Good navigation systems not only guide user to a destination, but also support them in understanding the environment, so they ultimately become able to reach the destination on their own. Understanding the environment also allows identifying alternative routes, e.g. shortcuts or along places worth seeing. This is only possible if users can acquire spatial knowledge about the environment when using a PND. Providing route instructions has, however, been shown to disengage the users from the environment [15] and make it difficult to understand the spatial layout of the environment [111].

Workload and Attention. According to theories and models about human cognition, such as the Multiple Resource Theory [30] or the PreNav model [26], the capacity of cognitive processing is a limiting factor when interacting with mobile devices. If a sense is already under high workload, it becomes difficult to process additional information through that sense. When walking through e.g. a crowded city centre the visual and auditory senses can be heavily occupied. Therefore, navigation information conveyed through visual and auditory displays might not be processed by the user. Studies with paper maps indicate that people can get very distracted by their navigation aid when navigating in unfamiliar environments [18]. Consequently, the users focus less attention on the environment. This is especially dangerous when passing through places with a high level of traffic, for example busy streets or crowded pedestrian zones.

Navigation Performance. Providing directions by navigation systems is usually quite effective in guiding travellers to their destinations. In particular, previous research indicates that people lose their orientation frequently when navigating in unfamiliar environments [11,18]. Recent studies suggest that people prefer paper maps over PNDs and navigate more efficiently with them [11,20]. Also, interacting with a mobile device is known to reduce the average walking speed [12], in general. Thus, navigating with a PND still offers room for improving the navigation performance.



Fig. 2. CoPilot for iPhone as an example for a pedestrian navigation system (<http://www.alk.com/copilot/> with courtesy of ALK Technologies)

4 Design of the *Tactile Wayfinder*

In this section we address the limitations of PNDs by advancing the concept of waypoint navigation with a tactile torso display. The basic idea is to convey the direction of the waypoint in relation with the user's heading. Previous groups have shown that this concept can effectively be used for waypoint navigation. [28][27][6][10]. In this work we employed this concept in a prototype called *Tactile Wayfinder*.

While being known that tactile waypoint navigation can reduce the cognitive workload, it conveys less spatial information than PNDs. Instead of showing a map that the user can use to learn how the route continues up ahead, the previously employed concept of tactile waypoint navigation just provides the location of the next waypoint. This reduced amount of spatial information might render it difficult to efficiently acquire spatial knowledge about the route. Thus, we investigated to enrich the spatial knowledge about the route presented by the tactile display. Our idea was to not only display the next waypoint, but the location of the subsequent waypoint as well. The subsequent waypoint then serves as a *look-ahead*, giving the user a cue about how the route will continue once the next waypoint has been reached (see Fig. 3). However, to realise this concept, we had to design a way of presenting the locations of the next waypoint and the look-ahead waypoint through different tactile cues.

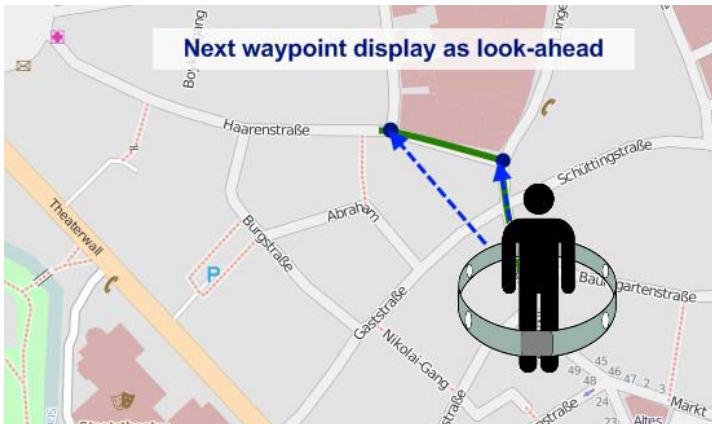


Fig. 3. Route visualisation through a tactile look-ahead: in alternating order the current waypoint and the subsequent waypoint are displayed. The user can anticipate how the route will continue beyond the current waypoint.

4.1 Tactons for Designing Tactile Cues

By introducing the notion of tactile icons (Tactons) Brewster and Brown [3] offered a concept to systematically design tactile cues. Tactons are abstract messages that can encode multidimensional information. In the case of the *Tactile*

Wayfinder we aimed at creating two two-dimensional Tactons where one information dimension encodes the direction of the waypoint and the other information dimension encodes the waypoint's type (next or look-ahead).

Six different tactile parameters can be used to encode information dimensions: amplitude, frequency, duration, waveform, rhythm, and body location. These parameters can then be combined to compose multidimensional messages where each parameter is mapped to one information dimension, e.g. the body location encodes the direction and rhythm encodes the type. Tactons composed of two or three different parameters can be encoded with a fairly high recognition rate of 70%, 81% [45]. However, in these studies the parameter space was limited to three levels of body location and rhythm, and two levels of waveform.

Which parameters can be used for Tacton design depends on the tactile actuators. For our work, we used a tactile belt with 12 vibration motors using off-centred weights to generate vibrotactile stimulations (see Fig. 4). These actuators are sewn into flexible fabrics, distributing themselves equally around the torso when worn. In order to indicate the location of a waypoint, the actuator which points most accurately into the waypoint's direction is activated. A built-in compass allows displaying absolute positions (e.g. North) independent from the user's orientation. Due to the off-centred weights the parameter space is limited. Changing the stimulus waveform is not possible with such actuators. Frequency and amplitude cannot be altered independently from each other, as they both depend on the applied voltage level, i.e. how fast the motor rotates. In this paper we refer to this combined parameter as *intensity*. This leaves us with four parameters for designing waypoint Tactons: intensity, duration, rhythm, and body location.



Fig. 4. The tactile belt we used for the *Tactile Wayfinder*

4.2 Design of the Waypoint Tactons

For presenting several waypoints to the user we needed to encode the waypoints direction and make them identifiable by encoding some form of waypoint type. Thus, we decided not to display distances in favour of a simpler Tacton design, since Veen et al. [28] have shown displaying the distance to a waypoint does not affect navigation performance.

As mapping body location to directions has shown to be intuitive and easy to understand [19][25] we incurred this concept. However, as we aimed at presenting two waypoints, we had to decide whether to present them simultaneously or successively. While presenting directions simultaneously has successfully been used [16] we decided against it, as distinguishing more than one waypoint in the same direction would not be possible. Instead, we chose to present the waypoints alternately.

Rhythm was used as parameter for encoding the waypoint type (next or look-ahead) since the study by Veen et al. [28] showed that different rhythms can be distinguished well when walking. In a set of informal tests we tested several rhythm patterns outdoors to ensure that the patterns could be easily identified when walking. The rhythm pattern that were finally used to encode the waypoint type are illustrated in Figure 5. The next waypoint is encoded by a heartbeat-like pulse which is repeated five times. The look-ahead waypoint is presented with a single pulse. Both Tactons are repeatedly presented with a duration of approximately four seconds per cycle.

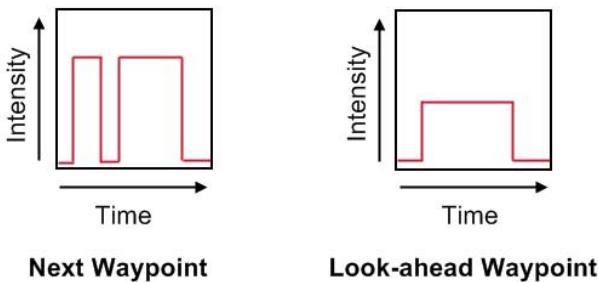


Fig. 5. The used Tactons: the heartbeat-like pulse (left) indicated the direction of the next waypoint. A single pulse (right) is used for the look-ahead waypoint.

4.3 Tactile Wayfinder Implementation

The tactile route visualisation with the waypoint look-ahead was integrated into the *Tactile Wayfinder* prototype. It was implemented using the open source Companion Platform for rapid mobile application prototyping [29]. As hardware platform we used a HTC Windows Mobile PDA. The belt was connected to the PDA via Bluetooth. A G-Rays 2 GPS receiver was used for obtaining the user's geo location. It was connected via Bluetooth as well.

5 Evaluation Method

To investigate if the concept of tactile waypoint navigation can overcome challenges of commercial PNDs we conducted an experimental field study. Participants had to use the *Tactile Wayfinder* and a commercial PND to reach a

destination in an urban environment (see Figure 6). The study took place on three consecutive Saturdays in May 2009. It took place in the city centre of Oldenburg. With its narrow, winding alleys the layout is rather complex and even residents sometimes have their problems in orienting themselves. Saturdays were chosen, as the city centre is most crowded on weekends. A complex layout and a crowded environment were suspected to increase the general cognitive load. In particular we investigated the effect of the navigation systems on the acquisition of spatial knowledge, the attention and cognitive workload, and the navigation performance. Our hypotheses were that:

- (H1) The *Tactile Wayfinder* will allow a better understanding of the environment in terms of landmark and survey knowledge than the PND,
- (H2) Users navigating with the *Tactile Wayfinder* will pay more attention to their environment compared to the PND, and
- (H3) The navigation performance of the *Tactile Wayfinder* will at least not be worse compared to the PND.



Fig. 6. A participant familiarises himself with the *Tactile Wayfinder*

5.1 Material

Two routes were created for the field study. Each route was about 800m long and contained six decision points. Both routes did not represent the shortest path to the destination and included awkward detours. Thus, good knowledge about the city centre was not privileged as the participants could not anticipate how the route would proceed beyond what the navigation system showed.

As PND, we chose TomTom¹ since it belongs to the state-of-the-art of pedestrian navigation systems. In a pilot test we investigated how to configure the PND most optimal. We found that map matching worked well in most cases.

¹ <http://www.tomtom.com/>

Sound was however turned off, as the pilot testers found it embarrassing and too hard to perceive. We configured TomTom to make use of the same type of bluetooth GPS receiver that the *Tactile Wayfinder* used. This ensured that the quality of the user position information was similar for both navigation systems.

5.2 Participants

Fourteen participants, seven female and seven male, took part in the study. The age ranged from 20 to 30 with a mean age of 25.33 (SD 4.51). In average, they rated their familiarity with the city centre to be slightly above average (2.71, SD 1.38 on a scale from 1=very good to 5=very bad). We also assessed their sense of direction through the SBSOD questionnaire by Hegarty et al. [9]. In average, our participants showed a neutral sense of direction (50.57, SD 17.62). However, there was a wide variability in the SBSOD scores. All participants signed an informed consent prior to the study. They were not paid for their participation.

5.3 Design

The navigation system served as independent variable. The *Tactile Wayfinder* represented the experimental condition while TomTom was used as control condition. The study used a within-subjects design. Thus, all participants contributed to both conditions. The order of conditions was counter-balanced to avoid sequence effects. The following dependent measures were taken in order to evaluate the acquisition of spatial knowledge, workload & attention, and the navigation performance:

Spatial Knowledge Acquisiton. The acquisition of spatial knowledge was measured by two tests that have been reported by Aslan et al. [1]. While the *photo recall test* is more focussed on landmark knowledge the *route drawing test* examines the survey knowledge. The *photo recall test* requires participant to recall how they turned at different decision points along route. These decision points are presented on photos and participants have to mark if they turned left, right, or went straight (if applicable). The score is taken by summarising the number of wrong answers. In the *route drawing test* the participants had to reproduce the route they just walked on a sheet of paper. As a reference, the sheet showed the starting point, the destination, and the outer bounds of the city centre. To determine the score for this test, we measured how accurate in terms of centimetres the waypoints of the route were drawn compared to a map of the city centre.

Workload and Attention. The level of attention was measured by assessing the subjective workload and counting how often the participants experienced near-accidents. Near-accidents were defined as situations where a participant nearly collided with another person or an obstacle. The participant had to be closer than 1 metre and perform a visible evasive manoeuvre. The subjective workload was assessed through self-report using the Nasa Task Load Index (TLX) [8].

Navigation Performance. The performance of the navigation task was measured in terms of completion time, disorientation time, and number of navigation errors. Disorientation events were counted when a participant explicitly mentioned to have lost orientation or when the participant stood still for more than 10 seconds. The event was considered ongoing until the participant continued to walk into the correct direction. Navigation errors were counted when the participants entered a street they were not supposed to. The completion time was the time it took the user to reach the destination.

5.4 Procedure

For each session the experimenters and the participants met near the starting point of the first route at a well known place. Before starting the actual evaluation, the participants had to fill out a questionnaire providing demographic information, judging their familiarity with the city centre, and answering the SBSOD items. The participants also learned that they had to complete spatial knowledge tests so they should pay attention to the route. The experimenters then explained the *Tactile Wayfinder* to the users and demonstrated the use of TomTom. The participants tested both devices before the measurements started. In alternating order, one of the navigation systems was then chosen for the first route. During the navigation task, the participants were asked to hold the GPS receiver in their hands during the evaluation. This was done to avoid the GPS signal being further distracted by being inside a pocket close to the body. Two experimenters followed the participants in some distance and noted near-accidents, navigation errors, and the number and length of disorientation events. When the participants arrived at the end of the first route they were asked to perform the two spatial knowledge tests (photo recall and route drawing) and rate the subjective workload. Then, the navigation system was changed and the participants started with the second route. Arriving at the second route's destination, the participants performed the spatial knowledge tests and filled out the Nasa TLX again.

6 Results

Spatial Knowledge. The score for the photo recall test was calculated by counting the number of wrong turning directions in the participants' responses. If the participants did not remember how they turned at an intersection shown on a photo we counted an error as well. If participants approached the decision point from an unexpected direction due to a previous navigation error, we compared the participants' answers to how they actually had turned. The results of the photo decision point recall test are shown in Figure 7. In average, TomTom users made 0.79 errors per route while *Tactile Wayfinder* users made 0.64 errors per route. There was no significant difference ($p = .34$).

Figure 8 shows one of the route drawings of the participants. The quality of these reproduced routes was quantified by comparing it with the actual route. We

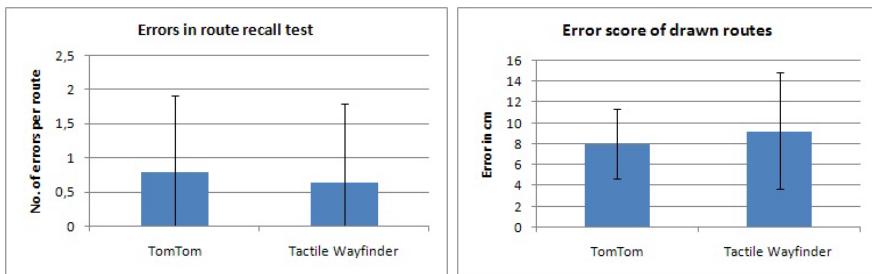


Fig. 7. The results of the spatial knowledge tests. The average errors in the photo recall test are shown on the left. The error score in cm from the participants' route drawing are shown on the right. In both tests there were no significant differences.

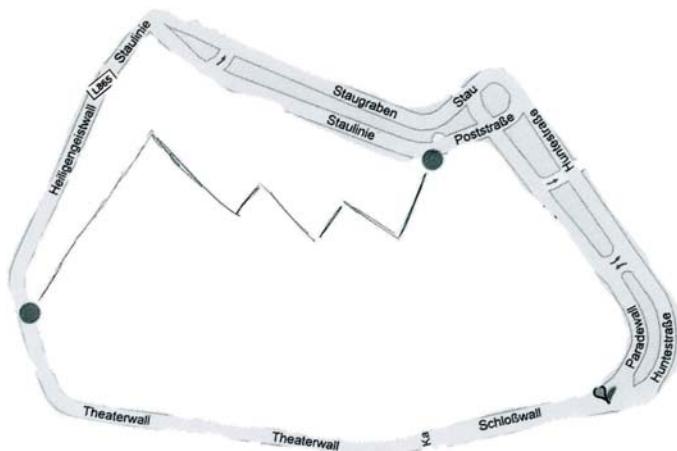


Fig. 8. One of the routes drawn by a participant after the evaluation. How accurate the participants could reproduce the routes were used to compare the spatial knowledge between the two conditions.

therefore scanned the drawings, printed them on transparent material, and put in on a map with the same scale. The distance in cm between each drawn waypoint and its correct counterpart served as error score. Figure 7 shows the average drawing error for both conditions. It was 8.02cm for TomTom and 9.25cm for the *Tactile Wayfinder* users 9.25cm. There was no significant difference ($p = .27$).

The participants' scores in the spatial knowledge tests had a small/medium correlation with the sense of direction SBSOD score ($r = -.21$ and $r = -.30$). This means that participants with a good sense of direction also had higher scores in both spatial knowledge tests.

Workload and Attention. Workload and attention were measured by self-report through the NasaTLX score and the number of near-accidents. Figure 9 shows

the average scores for both measures. As suggested in [8] we asked the participants to rate importance of each NasaTLX item for the navigation task. All possible pairs of items were presented and the participants had to choose the more important one. Mental demand and frustration were rated most important. Physical and temporal demands were rated least important. In average, TomTom users rated the workload with 2.78 and *Tactile Wayfinder* with 2.65. A higher score indicates a higher workload and seven was the highest score possible. There was no significant difference ($p = .40$). Figure 9 shows the number of near accidents. In average, TomTom users experienced 0.79 near-accidents/route while *Tactile Wayfinder* users had 0.14 near-accidents/route. Thus, significantly less near-accidents occurred with the *Tactile Wayfinder* ($p < .01$).

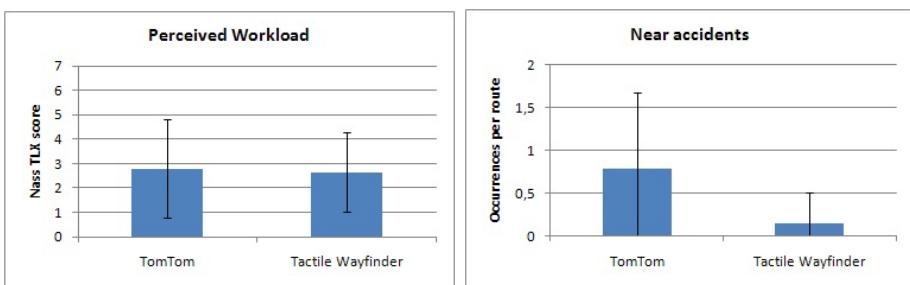


Fig. 9. Subjective workload rated by a Nasa TLX questionnaire (left). Number of near accidents (right).

There was also a noteworthy correlation between the sense of direction and the number of near-accidents. The seven participants with the lowest SBSOD score had 1.43 near accidents while those seven with the highest SBSOD only experienced 0.14 near accidents. Comparing the results of those groups statistically revealed a significant difference ($p < .001$).

Navigation Performance. The navigation performance was measured by the completion time, the number of navigation errors, and the time the participants were disoriented. Figure 10 shows the average results for both conditions. The average completion time was 763s/route with TomTom and 840s/route with the *Tactile Wayfinder*. There was no significant difference ($p = .09$). The number of navigation errors was 0.29/route for TomTom and 0.79/route for the *Tactile Wayfinder*. Participants with the *Tactile Wayfinder* made significantly more navigation errors ($p < .05$). The loss of orientation was measured in terms of how often and how long participants were disoriented. However, since both results are highly correlated ($r = .78$) we only report the disorientation time. The average disorientation time was 23.71s/route for TomTom and 36.00s/route for the *Tactile Wayfinder*. There was no significant effect ($p = .22$).

Completion time and loss of orientation count were highly correlated ($r = .77$). Thus, participants who often lost orientation were very likely to need more time

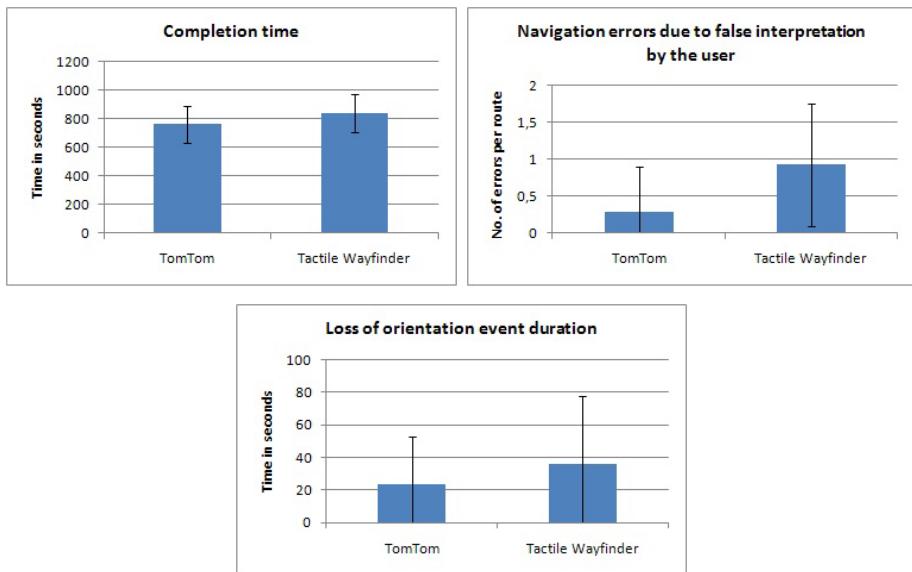


Fig. 10. The average performance measures for both conditions: completion time in seconds (left), navigation error count (right), disorientation time in seconds (bottom)

to complete the route. The number of navigation errors and the waypoint position errors while re-drawing the route were highly correlated ($r = .73$). Therefore, participants who made many navigation errors were also more likely to draw the walked route more inaccurately.

Gender Differences. In the experimental condition there were significant differences in the navigation performance between the genders. In average, female participants took longer to complete the routes ($p < .01$), made more navigation errors ($p < .05$), and lost their orientation more often ($p < .05$), as shown in table 1. However, this could only be observed with the *Tactile Wayfinder*. The navigation performance in the control condition was not significantly different.

Comments and Observations. Regarding TomTom the participants mainly concentrated on the route and their position shown on the map to navigate. To our surprise, none of the participants seemed to follow the turning instructions. This was a good idea since GPS was sometimes considerably inaccurate. It even

Table 1. The navigation performance with the *Tactile Wayfinder* split by gender

	Completion Time (s)	Navigation Errors	Disorientation Count
Female	925	2.43	3.29
Male	754.29	0.71	1.29

occurred that TomTom's map matching algorithm located people in the wrong street. Since the tactile belt employed a compass while TomTom depended on the GPS positioning update, there was a notable delay in updating the route. This turned into a problem for some of the participants as they turned into a new street but it took TomTom a few seconds to reflect that new situation. So, beyond each turning point there was a short period of "blind navigation".

Most errors with the tactile belt occurred at a y-formed junction of the second route where two paths continued almost in parallel direction. The tactile direction cueing combined with GPS inaccuracies was sometimes too coarse for the participants to clearly decide for one of the path. In some cases this caused the participants to choose the wrong path. This did not cause too much delay, since there was a connection between the two paths later on.

Regarding the subjective workload, the participants had divergent opinions: one half expressed that they found it exhausting to focus on the tactile cues. The other half stated that they could pay more attention to the environment. Many participants missed a map to get an overview about their environment. Some felt to be "bossed around" by the *Tactile Wayfinder*. Participants complained quite often that wearing the belt was uncomfortable due to the constant tactile feedback. Some suggested a tactile volume control or a pause function in order to be able to reduce the amount of feedback.

7 Discussion

In summary, both navigation aids enabled the participants to reach the given destinations. No difference in the participants' spatial knowledge acquisition could be found between the *Tactile Wayfinder* and TomTom. Using the *Tactile Wayfinder* the participants experienced fewer near-accidents but made more navigation errors. Having a better sense of direction correlated with fewer near-accidents and better spatial knowledge acquisition. Male participants had a better navigation performance with the *Tactile Wayfinder* than female participants.

Hypothesis H1 (the *Tactile Wayfinder* will allow a better understanding of the environment in terms of landmark and survey knowledge than the PND) could not be confirmed. Both spatial knowledge tests were insignificant. Thus, the tactile visualisation of the two upcoming waypoints could not improve the spatial knowledge acquisition compared to the PND. Instead, having a good sense of direction went along with better spatial knowledge scores. The sense of direction might therefore play a more important role in understanding the spatial layout of an environment than the actual navigation aid. Therefore, navigation aid designers should consider how to improve the general sense of direction along with the navigation instructions.

Hypothesis H2 (users navigating with the *Tactile Wayfinder* will pay more attention to their environment compared to the PND) could be confirmed. The *Tactile Wayfinder* allowed participants to spend significantly more attention to the environment so they were less likely to (nearly) collide with other people or obstacles. These findings confirm the predictions by Wickens' Multiple Resource Theory [31] or van Erp's Prenav model [26] that conveying information

via different senses reduces the overall cognitive workload. They also go conform with the results of Duistermaat et al. [6]. The subjective workload did however not decrease significantly. One explanation could be that the participants had to focus on the tactile output every now and then as there was no other source of directions. This goes along with the complaint that some participants felt to be "bossed around" by the *Tactile Wayfinder*. This could be countered by giving the participants a better overview of their situation, e.g. by combining tactile feedback with maps, as proposed in [22][18]. There was also a high correlation between a good sense of direction and few near-accidents. Thus, participants with a bad sense of direction paid less attention to the environment. It therefore seems important to specifically consider the group of users with a bad sense of direction in navigation system design.

Hypothesis H3 (navigation performance of the *Tactile Wayfinder* will at least not be worse compared to the PND) was refuted. The participants made significantly more navigation errors with the *Tactile Wayfinder*. The high correlation between completion time and disorientation events suggest that the participants lost most of their time when they were disoriented. On the other hand, this suggests that both navigation systems performed similar when the participants were well oriented.

The results also indicate that female participants had more problems navigating with the *Tactile Wayfinder*. Since the experimenter cannot assign the gender to the participants, gender-related results have to be analysed carefully. A simple explanation might be that the male participants were more tech-savvy in average. Another more interesting explanation can be found in the use of different wayfinding strategies reported in the literature. According to Lawton [14] women prefer a wayfinding strategy based on route-knowledge (e.g. at the shop turn left) while men prefer a survey-knowledge strategy (e.g. keep track of the own position on a map). Assuming our participants applied the respective wayfinding strategies, tactile waypoint navigation might not be compatible with route-knowledge-based strategies.

8 Conclusions

In this paper we investigated an approach to overcome existing limitations in commercial pedestrian navigation systems by advancing the concept of tactile waypoint navigation. To improve the spatial knowledge acquisition we conveyed two instead of a single waypoint at the same time. In a field study conducted in an urban environment, a commercial PND was compared with our *Tactile Wayfinder* in a navigation task. We could replicate previous findings that tactile information presentation can reduce cognitive load. On the other hand, the *Tactile Wayfinder* was outperformed by the commercial navigation system in terms of navigation errors.

The users' sense of direction turned out to be closely related to most of our results. A better sense of direction correlated with better spatial knowledge acquisition and a positive effect on the users' cognitive workload. In addition, a

better completion time was highly correlated with less disorientation events. The results also let us suggest that wayfinding based on survey knowledge (keeping track of the own location in relation to reference points) is correlated to a more successful navigation performance. Thus, being well oriented is important for spatial knowledge acquisition, cognitive workload, and navigation performance. Hence, future navigation systems should be designed to support the users' sense of orientation.

In summary trying to replace the traditional audio-visual interaction by tactile cues might not be the best idea. Instead we should look at how to combine the advantages of the tactile display's reduced required attention and the audio-visual systems' superior navigation performance. One fruitful approach therefore seems to be complementing those interactions, as e.g. proposed by [22][18].

Acknowledgments

The authors are grateful to the European Commission which co-funds the IP HaptiMap (FP7-ICT-224675). We like to thank Sren Samadi for preparing and conducting the study as well as our colleagues for sharing their ideas with us.

References

1. Aslan, I., Schwalm, M., Baus, J., Krüger, A., Schwartz, T.: Acquisition of spatial knowledge in location aware mobile pedestrian navigation systems. In: MobileHCI 2006: Proceedings of the 8th conference on Human-computer interaction with mobile devices and services, pp. 105–108. ACM, New York (2006)
2. Bosman, S., Groenendaal, B., Findlater, J.W., Visser, T., de Graaf, M., Markopoulos, P.: Gentleguide: An exploration of haptic output for indoors pedestrian guidance. In: MobileHCI 2003: Proceedings of the 5th conference on Human-computer interaction with mobile devices and services (2003)
3. Brewster, S., Brown, L.M.: Tactons: structured tactile messages for non-visual information display. In: AUIC 2004: Proceedings of the fifth conference on Australasian user interface, Darlinghurst, Australia, pp. 15–23. Australian Computer Society, Inc. (2004)
4. Brown, L.M., Brewster, S.A., Purchase, H.C.: A first investigation into the effectiveness of tactons. In: WHC 2005: Proceedings of the First Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, Washington, DC, USA, pp. 167–176. IEEE Computer Society, Los Alamitos (2005)
5. Brown, L.M., Brewster, S.A., Purchase, H.C.: Multidimensional tactons for non-visual information presentation in mobile devices. In: MobileHCI 2006: Proceedings of the 8th conference on Human-computer interaction with mobile devices and services, pp. 231–238. ACM, New York (2006)
6. Duistermaat, M., Elliot, L.R., van Erp, J.B.F., Redden, E.S.: Tactile land navigation for dismounted soldiers. In: Human factor issues in complex system performance, pp. 43–53. Shaker Publishing, Maastricht (2007)
7. Goodman, J., Gray, P., Khammampad, K., Brewster, S.: Using landmarks to support older people in navigation. In: Brewster, S., Dunlop, M.D. (eds.) Mobile HCI 2004. LNCS, vol. 3160, pp. 38–48. Springer, Heidelberg (2004)

8. Hart, S., Staveland, L.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Human Mental Workload, pp. 139–183. North Holland, Amsterdam (1988)
9. Hegarty, M., Richardson, A.E., Montello, D.R., Lovelace, K., Subbiah, I.: Development of a self-report measure of environmental spatial ability. *Intelligence* 30, 425–447 (2002)
10. Heuten, W., Henze, N., Boll, S., Pielot, M.: Tactile wayfinder: A non-visual support system for wayfinding. In: NordiCHI (2008)
11. Ishikawa, T., Fujiwara, H., Imai, O., Okabe, A.: Wayfinding with a gps-based mobile navigation system: A comparison with maps and direct experience. *Journal of Environmental Psychology* 28(1), 74–82 (2008)
12. Kane, S.K., Wobbrock, J.O., Smith, I.E.: Getting off the treadmill: evaluating walking user interfaces for mobile devices in public spaces. In: MobileHCI 2008: Proceedings of the 10th international conference on Human computer interaction with mobile devices and services, pp. 109–118. ACM, New York (2008)
13. Kray, C., Elting, C., Laakso, K., Coors, V.: Presenting route instructions on mobile devices. In: IUI 2003: Proceedings of the 8th international conference on Intelligent user interfaces, pp. 117–124. ACM, New York (2003)
14. Lawton, C.A.: Gender differences in way-finding strategies: Relationship to spatial ability and spatial anxiety. *Sex Roles* 30(11-12), 765–779 (1994)
15. Leshed, G., Velden, T., Rieger, O., Kot, B., Sengers, P.: In-car gps navigation: engagement with and disengagement from the environment. In: CHI 2008: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, pp. 1675–1684. ACM, New York (2008)
16. Lindeman, R.W., Sibert, J.L., Mendez-Mendez, E., Patil, S., Phifer, D.: Effectiveness of directional vibrotactile cuing on a building-clearing task. In: CHI 2005: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 271–280. ACM, New York (2005)
17. May, A.J., Ross, T., Bayer, S.H., Tarkiainen, M.J.: Pedestrian navigation aids: information requirements and design implications. *Personal and Ubiquitous Computing* 7(6), 331–338 (2003)
18. Pielot, M., Henze, N., Boll, S.: Supporting map-based wayfinding with tactile cues. In: MobileHCI 2009: Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 1–10. ACM, New York (2009)
19. Ross, D.A., Blasch, B.B.: Wearable interfaces for orientation and wayfinding. In: Assets 2000: Proceedings of the fourth international ACM conference on Assistive technologies, pp. 193–200. ACM, New York (2000)
20. Rukzio, E., Müller, M., Hardy, R.: Design, implementation and evaluation of a novel public display for pedestrian navigation: the rotating compass. In: CHI 2009: Proceedings of the 27th international conference on Human factors in computing systems, pp. 113–122. ACM, New York (2009)
21. Rukzio, E., Schmidt, A., Krüger, A.: The rotating compass: a novel interaction technique for mobile navigation. In: CHI 2005: CHI 2005 extended abstracts on Human factors in computing systems, pp. 1761–1764. ACM, New York (2005)
22. Smets, N.J.J.M., te Brake, G.M., Neerincx, M.A., Lindenbergh, J.: Effects of mobile map orientation and tactile feedback on navigation speed and situation awareness. In: MobileHCI 2008: Proceedings of the 10th international conference on Human computer interaction with mobile devices and services, pp. 73–80. ACM, New York (2008)

23. Tan, H.Z., Pentland, A.: Tactual displays for wearable computing. In: ISWC 1997: Proceedings of the 1st IEEE International Symposium on Wearable Computers, Washington, DC, USA, p. 84. IEEE Computer Society, Los Alamitos (1997)
24. Tsukada, K., Yasumura, M.: Activebelt: Belt-type wearable tactile display for directional navigation. In: Davies, N., Mynatt, E.D., Siiro, I. (eds.) UbiComp 2004. LNCS, vol. 3205, pp. 384–399. Springer, Heidelberg (2004)
25. van Erp, J.B.F.: Presenting directions with a vibrotactile torso display. Ergonomics 48, 302–313 (2005)
26. van Erp, J.B.F.: Tactile Displays for Navigation and Orientation: Perception and Behavior. PhD thesis, Leiden, The Netherlands (2007)
27. van Erp, J.B.F., van Veen, H.A.H.C., Jansen, C., Dobbins, T.: Waypoint navigation with a vibrotactile waist belt. ACM Trans. Appl. Percept. 2(2), 106–117 (2005)
28. van Veen, H.A.H.C., Spap, M., Van Erp, J.B.F.: Waypoint navigation on land: Different ways of coding distance to the next waypoint. In: Proceedings of the 4th International Conference EuroHaptics 2004 (2004)
29. Wichmann, D., Pielot, M., Boll, S.: Companion platform - modular software platform for rapid development of mobile applications. It - Information Technology 51(2), 72–78 (2009)
30. Wickens, C.: Multiple resources and performance prediction. Theoretical Issues in Ergonomics Science 3(2), 159–177 (2002)
31. Wickens, C.D.: Processing resources in attention. In: Processing resource in attention. Academic, London (1984)

Jog Falls: A Pervasive Healthcare Platform for Diabetes Management

Lama Nachman¹, Amit Baxi², Sangeeta Bhattacharya², Vivek Darera²,
Piyush Deshpande², Nagaraju Kodalapura², Vincent Mageshkumar², Satish Rath²,
Junaith Shahabdeen¹, and Raviraja Acharya³

¹ Future Technology Research, Intel Labs, Santa Clara, CA

² Future Technology Research, Intel Labs, Bangalore, India

³ Kasturba Medical College, Manipal University, Manipal, India

Abstract. This paper presents Jog Falls, an end to end system to manage diabetes that blends activity and energy expenditure monitoring, diet-logging, and analysis of health data for patients and physicians. It describes the architectural details, sensing modalities, user interface and the physician's backend portal. We show that the body wearable sensors accurately estimate the energy expenditure across a varied set of active and sedentary states through the fusion of heart rate and accelerometer data. The GUI ensures continuous engagement with the patient by showing the activity goals, current and past activity states and dietary records along with its nutritional values. The system also provides a comprehensive and unbiased view of the patient's activity and food intake trends to the physician, hence increasing his/her effectiveness in coaching the patient. We conducted a user study using Jog Falls at Manipal University, a leading medical school in India. The study involved 15 participants, who used the system for 63 days. The results indicate a strong positive correlation between weight reduction and hours of use of the system.

Keywords: Personal Health Monitoring, Diabetes Management, Energy Expenditure Analysis, Activity monitoring.

1 Introduction

Metabolic syndrome is emerging as a major public health issue across the world. It is a group of symptoms (e.g. central obesity, high blood pressure, insulin resistance) that increase the risk of heart disease, type-2 diabetes and stroke. Healthy life style characterized by increased physical activity and moderation of eating habits plays a key role in reducing these risk factors and slowing down the progression of type-2 diabetes or possibly even reversing it.

To empower patients to better manage their disease, we envision a system that allows them to continuously monitor their physical activity and food intake, set goals and monitor progress towards these goals, allow them to reflect on their trends over an extended period of time and draw some actionable conclusions. Since physicians play an important role in coaching the patients, this system should be an effective tool

in enabling physicians to be better coaches. Specifically we envision a backend portal giving physicians comprehensive and unbiased visibility into the patients' life styles with respect to activity and food intake, as well as enabling them to track their progress towards agreed upon goals. To accomplish this vision, we worked closely with physicians to design a system for diabetes management which we call Jog Falls.

In this paper we describe our Jog Falls system, its high level architecture and detailed design, highlighting the novel contributions in the different tiers of this architecture. Jog Falls is an integrated system for diabetes management providing the patients with continuous awareness of their diet and exercise, automatic capture of physical activity and energy expenditure, simple interface for food logging, ability to set and monitor goals and reflect on longer term trends. Its backend interface enables the physician to view the progress and compliance of the patients, hence facilitating personalized coaching. Finally its novel method for fusing Heart Rate (HR) and accelerometer data improves the accuracy of energy expenditure estimation, a key feature in enabling weight loss. We conducted a user study using Jog Falls at a leading medical school in India. India is on the verge of being the diabetes capital of the world [1], which calls for a comprehensive study in the Indian context. The study involved 15 participants, who used the system for a period of 63 days. We report the results of the study and discuss the effectiveness of this system in helping patients manage their lifestyles. Specifically the results indicate a strong positive correlation between weight reduction and hours of use of the system.

The rest of the paper is organized as follows. We first present an overview of related work in section 2. We then describe the system design and implementation in section 3 and follow up with the details of our user study and the key learnings in section 4. We conclude with section 5.

2 Related Work

Currently available solutions for diabetes management are discrete and disconnected. Activity and calorie expenditure monitoring devices such as pedometers quantify activity in terms of "number of steps" walked or calories expended. They use either spring based mechanical sensors [2], piezoelectrics [3] or accelerometers [4], [5] for activity monitoring. These devices are comparatively inexpensive and provide a reasonable solution for quantifying limited states like walking and running. However, they are not suitable for comprehensive activity tracking and are also susceptible to vibration errors. BodyBugg [6] uses additional sensors like heat flux, skin conductivity and temperature along with an accelerometer to improve the accuracy of calorie estimation. Several studies have positively evaluated the validity of BodyBugg in laboratory conditions [6]. The device is harnessed using an armband, which makes it a user friendly solution for long term monitoring. However, the lack of a user friendly way to track and set calorie goals is a major drawback of this system. Polar [7] uses Heart Rate (HR) to quantify activity in terms of calories, and is ideal for sports and fitness training applications. However, it is not designed for tracking continuous energy expenditure over extended time and doesn't compensate for psychological factors that affect HR in sedentary conditions. Moreover, this device uses a generic calibration curve (HR vs MET value) for all users, which reduces its accuracy.

UbiFit [8] is a system developed to encourage individuals to self-monitor their physical activity and incorporate regular and varied activity into everyday life. The system components like glanceable display, interactive application, and a fitness device [9] provides valuable features like instant feedback on goals, long term log and journaling of activities and automated inference of several day to day activities. However, the system does not estimate calories burned while performing different activities. Other solutions like Houston [10], SHAKRA [11] and Fish'n'Steps [12] also suffer from the same limitation. Houston and Fish'n'Steps use pedometers to track step counts while SHAKRA uses cell phone signal strength variation patterns to track user activity. In Houston and SHAKRA, users can view real-time and historical step count/activity on the phone and also share this information with a group of friends. In Fish'n'Steps, user activity is presented as a game designed to encourage behavior change via social co-operation and competition.

For dietary tracking, a major gap in today's systems is keeping track of calories consumed in real-time. Websites like Nutrition Vista [13], Body Media [14], and FitDay [15] provide calorie intake estimates when the user enters the diet consumed, making them susceptible to recollection errors and hindering real-time self awareness. Myfoodphone [16] tracks food intake using photographs of the food captured by the users, thereby creating a photo-journal. The photos are sent using GSM/GPRS to a dietician, who assigns a goodness score based on preset targets. It also incorporates social facilitation, enabling the user to see progress of others. This system lacks real-time, automated feedback and requires regular intervention from the nutritionist. An iPhone app from Weight Watchers [18] features a large menu of food options, a point calculator that gives a score to users based on diet, and recipes that can be used to meet the point targets of the users.

MAHI [17] is a solution to help newly diagnosed diabetics develop reflective thinking skills through social interaction with diabetes educators. It supports capture of interesting events that "disrupt regular activities" as opposed to capturing pre-defined activities like meals and exercise. While MAHI provides the user the flexibility to log experiences using voice notes and photographs as well as consult with diabetes educators, it does not support real-time activity and calorie expenditure/intake monitoring.

It is clear that there is no solution that encompasses all the above elements in an integrated, simple and usable manner. Other features not readily available include enabling physicians to set goals remotely, regularly monitoring the patient's condition, his/her adherence to food and activity goals, and coaching and motivating the patient to modify his/her lifestyle on a continuous basis. We tried to address these gaps by providing an integrated, usable and comprehensive framework that encompasses and manages all the above elements of diabetes management.

3 System Design and Implementation

We defined the requirements of our Jog Falls system by working closely with physicians. The main goal of the system is to empower patients to manage their life style with respect to diet and exercise hence lowering their risk factors for metabolic

syndrome, and enable physicians to be more effective in helping their patients reach their goals. To accomplish these objectives, the following requirements were defined:

- The system needs to enable patients to monitor their diet and exercise and give them continuous awareness of caloric intake and expenditure. Since a change in behavior is required, patients need to first understand how different activities and diet choices impact their calorie expenditure and intake, and how these two factors can be balanced to achieve their final goal.
- The system should enable the physician to help the patients set and reach these goals. To be effective, the physician needs a comprehensive and objective visibility into the patients' diet and exercise, instead of relying on their recollection, which may be incomplete, inaccurate or both. The physician needs access to this data on weekly basis without any user intervention.
- The system needs to continuously log physical activity along with accurate (within 80%) translation to energy expenditure. Due to the comprehensive nature of activity logging (all daily activities), this logging needs to be automated.
- Since food intake is intermittent, it is acceptable to have the users manually enter what they consume, but the system needs to calculate the resulting calories based on food choices. The food options need to be customized to Indian diet.
- The system needs to be available for at least 15 hours per day (to cover the “waking hours”) and should only require charging once per day (overnight)
- The system should enable goal setting for energy expenditure and food intake as well as track and inform the user of the progress. Real-time positive feedback as well as intervention when not meeting the goals are necessary.



Fig. 1. Jog Falls is a three tier system for diabetes and metabolic syndrome management

3.1 System Overview

To satisfy the requirements of this application, we designed Jog Falls, an end to end solution for management of diabetes, pre-diabetes and metabolic syndrome. Jog Falls has a 3 tier architecture as shown in **Fig. 1**. The first tier consists of the sensor devices

responsible for collecting the physiological and activity data, and as a result need to be continuously worn on the body. The second tier consists of a smart phone, which is responsible for communicating with the sensors via bluetooth, aggregating and storing the sensor data, calculating the energy expenditure and intake, providing the user interface for logging, alarming and data review, and communicating with the third tier through GPRS. GPRS was chosen due to its availability and the relative lack of broadband or landlines in India. The third tier consists of a backend server that is responsible for aggregating and storing the data from all users, and providing the user interface for the physician.

3.2 Tier 1: Sensing Components

To provide accurate estimation of Energy Expenditure (EE) i.e calories burned, we chose to fuse accelerometer and heart rate data. Heart Rate (HR) can be used to measure EE, since EE is fairly linearly proportional to sub-maximal Heart Rate (HR). However, HR can be influenced by psychological & emotional factors, drugs and caffeine which will result in inaccurate EE estimation. Accelerometer based sensors are used to estimate the current physical activity [19] which can be mapped to energy expenditure, and are not affected by emotions or drugs. However, they can only recognize a limited set of activities that they were trained for, and are not able to accurately estimate effort. For example, it won't be able to distinguish between walking and walking with a heavy backpack . Previous studies show that combining accelerometer data from chest and hip provides better estimation of physical activity in comparison to using single accelerometer [20]. Jog Falls uses two body-wearable sensors (HR-SHIMMER and MSP) for continuous monitoring of a user's activity throughout the day. A data fusion algorithm uses Heart Rate (HR) and accelerometer data to improve the EE accuracy by addressing the limitations of these individual modalities. Apart from these sensors, off-the-shelf sensors were used to keep track of user's blood pressure (twice a day) and weight (once a week).

3.2.1 Mobile Sensing Platform

The Mobile Sensing Platform (MSP) [22], shown in Fig. 2 (c), is a battery operated device equipped with multiple sensors like, 3D accelerometer, light, barometer, humidity, microphone and temperature. The platform supports many applications, like inertial navigation, and user activity inference and can be worn on the waist. We developed a hierarchical adaptive boosting classifier that uses the accelerometer data to discern states like sitting, standing, laying, strolling, brisk walking and running. The hierarchy allowed us to make better use of the feature space in accurately estimating both sedentary and active states. The classified decision vector containing the most probable user activity and speed of the user is sent to the aggregator every 5 seconds to facilitate further processing. The inferred activity is also used directly by the aggregator for real time user feedback and trend analysis. Since these users were expected to be sedentary quite often, further classification within the sedentary states was important to accurately estimate total energy expenditure.

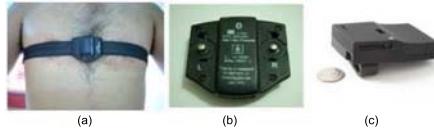


Fig. 2. Sensors for monitoring heart rate (a,b) and physical activity monitoring (c)

3.2.2 HR-SHIMMER

HR-SHIMMER, shown in Fig. 2 (b) is a battery operated device, which can be worn on the chest by means of a conductive fabric chest belt (Fig. 2 (a)), and is used to estimate the *intensity* of activity and Energy Expenditure (EE). This device, based on the SHIMMER platform [23], has an ECG front-end to measure Heart Rate (HR), a 3-axis accelerometer to measure upper body movements, a microcontroller and a Bluetooth interface. Some of its key features are:

- *Quantification of upper body movement:* Energy spent in low intensity sedentary activities (e.g. working on PC, bending, desk work, cooking, ironing) forms a major part of the Total Daily Energy Expenditure (TDEE), making accurate EE estimation of such activities quite important. To address this requirement, HR-SHIMMER quantifies the intensity of upper body movements using its internal accelerometer. The frequency and magnitude of the resultant signal from the three axes are used to estimate the intensity of even the most subtle body movements in terms of a derived parameter called Sedentary Metabolic Equivalents (S-METs). This enables accurate quantification of low intensity activities (less than 3 METs) usually carried out while sitting or standing. Since the effect of psychological factors on HR is more pronounced when the user is sedentary, using S-MET (instead of HR) for quantification of sedentary activities reduces the effect of such factors. For activities with intensity equal to or greater than slow walk (S-MET greater than 3 METs), the EE is estimated from HR. This helps prevent EE overestimation while traveling in vehicles.
- *Off-body Detection:* In order to understand the device usage patterns it is essential to know when and for what duration, the device is worn on the body. To enable this requirement, an algorithm on HR-SHIMMER analyses the ECG baseline stability, presence of noise and absence of ECG signal to detect when the device is off-the-body. During off-body condition, HR calculation is disabled and the device sends “off-body” status to the Aggregator.
- *Individualized HR calibration:* For estimating EE, most commercial heart rate monitors assume a common HR calibration curve for all individuals. This approach results in EE estimation errors since the change in HR of an individual when subjected to an exercise load depends on the fitness level. Hence it is necessary to calibrate HR with respect to exercise load (which is quantified in terms of METs), for each individual. Current HR calibration techniques based on Oxygen Uptake (VO_2) require specialized equipment like oxygen and carbon dioxide gas analyzers, treadmill in a lab setting [24][25], which limits their use. In order to enable HR calibration in the field, we developed a novel simplified HR calibration technique for EE estimation from HR with a comparable level of accuracy. Our calibration procedure involves measuring the HR at rest and at different exercise loads (slow

walk, medium walk, brisk walk and jog). The walking speed is calculated by measuring the time required to cover a known distance. An application on the Aggregator device computes the METs for each walking speed (using published MET values [26]), the corresponding HR, and then builds an individualized HR v/s METs calibration curve. Fig. 3 shows calibration curves for two users using our calibration method compared to the VO_2 calibration method. Fig. 5 shows the EE estimation accuracy using this calibration technique relative to EE estimated using VO_2 . Re-calibration may be required if the fitness level changes significantly.

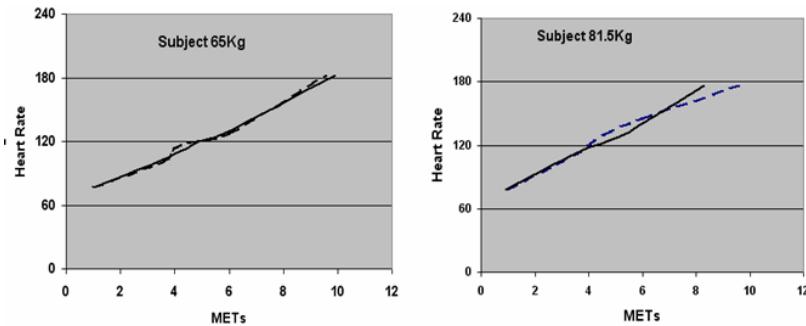


Fig. 3. Comparison of our calibration technique with VO_2 method. Dashed line is based on our Calibration algorithm and continuous line is based on VO_2 method.

3.2.3 HR + Accelerometer (HR+A) Data Fusion

Estimation of EE in an accurate and reliable manner was a firm requirement of our application. Current techniques for estimating EE in ambulatory condition involve sensors like heart rate, accelerometer, body temperature and galvanic skin response used either individually or in combination. However, existing systems fail to address low intensity upper body activities, psychological effects on heart rate in sedentary states and data losses from the sensors. In addition the existing approaches use a single accelerometer [27], which reduces the accuracy of physical activity estimation [28], in turn affecting EE estimation. To address the above limitations, we developed a data fusion algorithm (HR+A), shown in Fig. 4, which fuses HR with accelerometer data from HR-SHIMMER (upper body) and MSP (lower body). Our algorithm (Fig. 4) running on the aggregator, chooses the correct sensor to calculate the metabolic equivalent value using features like offset, forward difference, resting heart rate, together with lower and upper body movement intensity from the accelerometers.

The algorithm was based on a lab experiment that we conducted in the Human Performance Laboratory of Oregon Health Science University. Four subjects wearing a hip mounted accelerometer, Body Bug [6], and Polar HR monitor [7] were also connected to a Calorimeter. Required parameters like, lower / upper body movement, heart rate, and EE from the calorimeter were logged. The data from the experiment was used to train a Bayesian network to choose the sensor for EE calculation that will maximize proximity to the Calorimeter. Cases like exercise recovery, upper body workout and exertion effect were captured accurately in the training. Other effects like emotions and moving vehicles were added later, due to the difficulty of collecting

such data in a lab setting. The experiment data was split in two, where the data from two users was used for training the Bayesian network and the data from other two was used to validate the network. Graphs comparing the EE estimated using our algorithm, and EE estimated from calorimeter are shown in Fig. 5.

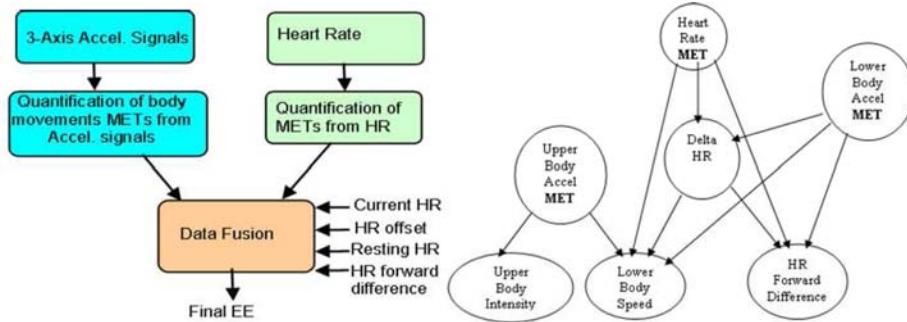


Fig. 4. Data Fusion algorithm and comparison with VO_2 method

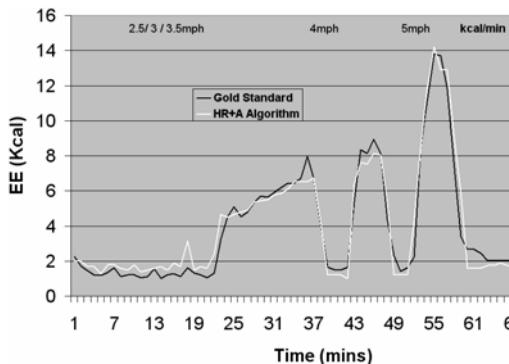


Fig. 5. Comparison of our EE estimation with VO_2 method

The EE estimation accuracy was also tested in free living conditions with BodyBugg [6], a state-of-the-art multi-sensor device, the results of which are shown in Table 1. The ability to perform partial inference using Bayesian network allowed us to operate with nominal accuracy even with missing data. This feature proved essential given the sensor data loss in the field due to battery depletion or user error.

Table 1. Comparison of EE accuracy of HR+A algorithm with BodyBugg

Test	HR+A(kcal)	Bodybugg(kcal)	% difference
Home Activities	319	340	-6%
Working at Desk	128	125	2%
Full Day Test	1308	1378	-5%

3.3 Tier 2: Aggregator

We chose the HTC Touch smart phone (201 MHz processor, 128 MB RAM, 1800mAh battery pack, Microsoft Windows Mobile 6 OS) as the aggregator and built our application on top of a custom framework that provided sensor communication, data storage and data synchronization with the backend and the ability to plug in different analysis and UI modules to customize the application. We used Merge Replication (a feature provided by MSSQL server) over GPRS to synchronize the data with the backend server. GPRS was chosen since we found that Wi-fi/WIMAX has minimal penetration among the users of our evaluation. Data would be sent over the GPRS network every night, between 11pm to 6am while the device was charging to minimize effect on battery life. If the synchronization process failed consecutively on 3 nights, the application would try to send the data during the day, at the risk of stopping data collection from the sensors. Some of its key features are:

Self Awareness: To provide users continuous self awareness of their exercise and diet, we designed the MyDay screen as shown in Fig. 6(a). It is a responsive glanceable display to encourage users to view on the go their Energy Expenditure (EE), Calorie Intake (CI), Heart Rate (HR) and activity type and intensity. The HR of the user is shown in the top left of the screen. This is updated every 5 seconds. The 2 status bars show the calorie intake and expenditure for the current day. The calorie expenditure is updated on this screen every minute. The status bars show how the user is tracking to the ‘target’ that has been set by the physician, based on the user’s BMI and medical history. If the user is ‘trailing’ the target at a given time in the day, the numbers turn RED in color.

A turtle avatar (a turtle symbolizes luck and the fact that ‘slow and steady wins the race’) was chosen to showcase different states, according to the current activity of the user (as inferred by the MSP activity sensor). The state of the turtle is updated every 5 seconds. The different states of the turtle are shown in Fig. 6(b). It also provides recommendations to the user based on user state, and the state and color of the turtle indicate to the user how well he is doing at a given point in the day with respect to his activity goals. If the user has been sedentary (sitting or lying down) for more than ‘x’ minutes over the past ‘y’ hours (x and y can be set by the physician), the shell of the turtle turns RED to indicate that action is required by the user, and stays RED as long as the condition is true. There is also a dialog box next to the turtle where custom messages can be displayed to motivate users to do some form of physical activity on regular basis.

Trends: Self-awareness is also supported through trend information. This provides opportunities for the user to notice patterns of success and failure, and hence improve in the long term. Weekly and daily views of calorie expenditure, calorie intake, activity intensity and activity type (Fig. 7) are provided. It can also show past measurements of Blood Pressure, Weight and Blood Glucose (first graph in Fig. 7).

Goal Setting: Since the system is intended for long term use, it was desired to enable tweaking goals to challenge users, in accordance with behavioral theories like the Goal Setting Theory [29]. Based on body weight, BMI and physical activity, the desired target weight loss is calculated by the physician. This target is then mapped to food intake and activity goals based on discussions with the user and possible consultation with a nutritionist. It is important to enable readjusting these targets over time

based on the user's performance. Since goal setting is envisioned to be a shared responsibility between patient and physician, the system allows for remote setting of these goals from the backend tool (calorie intake and expenditure), which are transmitted to the aggregator and refreshed on the subsequent day.

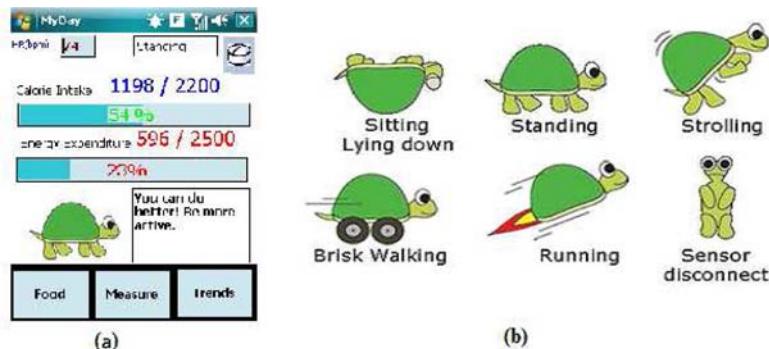


Fig. 6. MyDay screen on the Aggregator (a) and different activities depicted by the turtle (b)

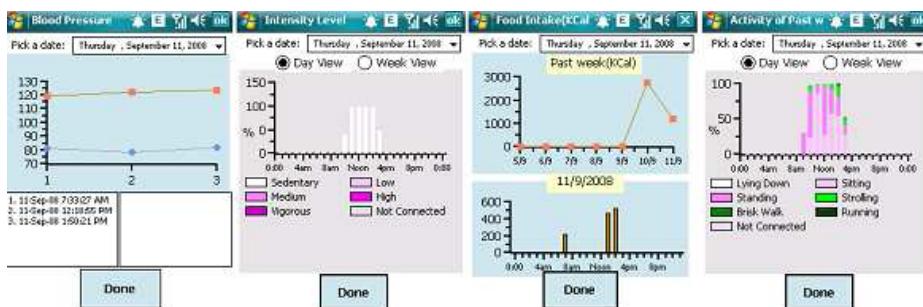


Fig. 7. Example trend graphs

Diet Logging: Users can log their diet intake using photograph method and/or item entry (Fig. 8). In the photograph method, the user can snap a picture of the meal, which is time-stamped and stored in the phone, and later transmitted to Tier 3. This method requires minimal user input, but the user does not get real-time feedback of the calories consumed. The phone contains a comprehensive database of Indian food items along with their caloric values, and a picture of that item. The calorie intake is calculated based on this information. The food item entry has been designed for easy and quick entry. The default food menu depends on time of day simplifying data entry. There is also a provision for alphabetical search. Once the user finds the desired food item, he is required to enter the quantity and the size of the item that has been consumed. There are 3 options for size entry: small, medium and large. The picture of the food item corresponds to a 'medium' portion of the food item, to help the user understand that the meaning of each subjective term.

Most Indian foods are usually not eaten in isolation, but with an accompanying gravy/vegetable. To simplify data entry, a list of associated food items (up to 3) appear in the same screen as the main entry. The user is also shown the calories of the food items that he has selected to facilitate learning. After all food items have been entered, the user is shown all selected items in a review screen for confirmation.

Finally, the user can enter weight, blood sugar and blood pressure values and look at trends of these values over a period of time. The system also supports setting of reminders for medicine intake. A reminder is displayed on the top-right corner of the MyDay screen as seen in Fig. 6(a).



Fig. 8. User interface for food logging

Behavior Change Strategy: Jog Falls employs several persuasion techniques like goal-setting, positive-feedback, comprehensive-behavior-coverage [29], reduction, tunneling, tailoring, self-monitoring, surveillance, conditioning [30] and just-in-time suggestions [30][31] to motivate behavior change. It supports self-monitoring via simple interfaces (reduction). Goal-setting, surveillance (external observation) and tunneling (guided persuasion) are achieved via the involvement of a physician. Tailored suggestions like “You have been sedentary for too long! Please get active!” are provided at the right time to encourage physical activity. Furthermore, positive reinforcements (conditioning) like “Great Job, Keep Going!!” are used to encourage desirable behavior. Lastly, Jog Falls supports energy expenditure computation for all activities (comprehensive-behavior-coverage) hence not limiting the user’s activities.

3.4 Tier 3: Backend Application

Dia-Graph is the backend application software running on a web server and can be securely accessed from remote client machines over IP network. The Dia-Graph application is designed to enable the physician to study the patient’s life style and provide necessary advice/coaching to better manage chronic disease conditions. In order to fulfill these requirements, Dia-Graph offers the following key features:

- *Energy expenditure and calorie intake mapping:* One of the challenges in managing diabetes is to keep track of one’s energy expenditure and calorie intake on a continuous basis. Dia-Graph bridges this gap by providing a time synchronized EE

and Calorie Intake mapping display to enable the physician to track the progress of both diet and energy expenditure in a simplified manner.

- **Activity distribution:** Dia-Graph can represent the activity distribution over an extended duration. It supports two types – activity type and activity intensity graphs. Using these graphs, the physician gains insight into the type of activity the user is performing and the times of the day the user is most active and provide coaching accordingly. Energy expenditure is also displayed facilitating user education with regards to the impact of different activities.
- **Target and reminder settings:** Another key feature of Dia-Graph is to enable the physician to remotely set goals/targets for the users based on activity and diet, track their progress and provide coaching tips to help motivate them. The physician could also set reminders (e.g. medicine intake) using the Dia-Graph.
- **Consolidated report generation:** Dia-Graph allows the physician to automatically generate a consolidated report on various user parameters and print out the report.



Fig. 9. Dia-graph application for the physician

4 Evaluation Study

To evaluate the effectiveness of Jog Falls against the stated goals, we conducted a pilot study at Manipal University in collaboration with the co-PI Dr Acharya. The trial involved 15 participants, who evaluated the system for a period of 63 days.

4.1 Study Design and Participants

Participants were selected based on a certain set of inclusion and exclusion criteria. Some of the inclusion criteria included (a) adults between 18 to 60 years of age, (b) family history of diabetes, (c) impaired glucose tolerance, (d) high risk of type 2 diabetes mellitus (e.g. obese/overweight) and (e) willingness to participate with a basic aptitude to handle a cell phone. No discrimination was made based on sex or profession. The only exclusions involved people with serious medical illnesses and pregnant women. Out of the final 15 participants selected, 3 had diabetes and 11 were at risk due to one or more of the following factors - overweight, obesity, family history, increased lipids or hypertension; 1 participant was female and the rest were male. At the start of the study, participants were given a demonstration of the system and a short training. They were given an HTC touch phone with installed software, an MSP sensor and a HR-SHIMMER sensor with polar chest belt. Furthermore, they were instructed to:

- Use the system throughout the day, between 6am and 9pm and charge the phone and sensors during night hours.
- Not use the system when swimming, having a bath and when traveling by air.
- Refer to the provided troubleshooting instructions and user manual and/or contact co-PI. They were asked to maintain a logbook of problems faced and comments.
- Meet with the co-PI every Saturday, to discuss issues, track their progress and provide feedback about the system. During these visits the co-PI analyzed the weekly data of each participant and provided appropriate recommendations.

Fixed calorie intake and expenditure goals were set per participant, by the co-PI, to achieve a 5% weight reduction over the duration of the study. Personal details of participants like identification data, demographic data, education, socioeconomics, technology background, behavioral attributes and relevant background health information were collected at the start of the study. Participant feedback was collected at the end of the first week, during visits with the co-PI and at the end of the study.

4.2 Study Results

At the end of the study, we analyzed participant feedback and system usage data to evaluate whether JogFalls operated as expected, was useful to the participants and the physician, and whether it influenced participant health behaviors.

4.2.1 Effectiveness of the System in Weight Reduction

Overall, the results give convincing statistical evidence on actual clinical usefulness of the system. There was significant mean 0.85 ± 1.68 kg ($0.72 \pm 1.52\%$ of target) and median weight loss 0.99kg among subjects at the end of the trial. We used historical control wherein weight records of the patients with similar clinical profile and weight reduction needs were collected from the case sheets. Mean weight change of such nine subsequent subjects with components of metabolic syndrome in medical OPD of Manipal University who received standard care and advice on weight reduction (without device but with bi-weekly or monthly sessions with physicians) and life style measures had mean weight gain of 0.33 kilograms over two months period. The other

striking outcome from the analysis of the intensity of the use was very strong positive correlation (P value 0.001) between weight reduction and hours of use by subjects, as shown in Fig. 10(a).

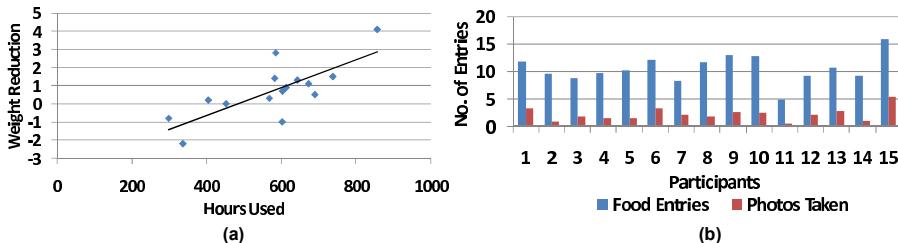


Fig. 10. (a) Weight loss vs system usage and (b) Average food entries and pictures per day

4.2.2 Usability

We analyzed the usage patterns to determine system usability. We found, that with the inclusion of non-usage of system due to technical reasons, travel by subjects, etc., the system was substantially used (85% of the days, and 12.7 hrs per day on average).

Participant feedback showed that overall, most participants found the system easy to use. A few participants faced problems with the chest belt, requiring either readjusting the belt, or remoistening of the lead area of the chest belt, or changing belts during the day to avoid discomfort due to excessive sweating. Some participants also developed rashes due to the belt which later subsided with medication and regular washing of the belt. Overall, 8 participants found wearing the chest belt easy, 6 found it to be manageable and 1 found it to be difficult. Participants also preferred to carry fewer devices and suggested that the number of devices be reduced from 3 to either 1 or 2. Regarding the MSP, two participants faced problems due to the MSP falling off the waist and 5 participants had broken MSP clips. Participants also found the blinking lights on the MSP uncomfortable, since it tended to draw undue attention. 5 participants expressed that they would like the MSP to be more rounded and less bulky, with reduced vertical height to increase comfort. Additionally, participants expressed a desire for longer battery life of the sensors.

We also analyzed the logged data in the backend and on the phone to evaluate the usage of some of the application features. The analysis revealed that the food logging feature was used on 93% of the days and that people had made an average of 10 food entries per day. This large number of per day food entries is expected given the numerous items taken per Indian meal. We also found that people had taken about 2.24 pictures per day. This is close to the expectation of 3 pictures per day for the 3 main meals of the day. During the weekly visits, we observed that the co-PI would correlate the pictures with the food logging entry and, if required, correct the notion of the ‘small’, ‘medium’ and ‘large’ portions. Thus, over a period of a few visits, users were calibrated to the meaning of these terms. The distribution of the average number of per day food entries and photos taken is shown in Fig. 10(b). We believe that this data indicates ease of use and acceptance of these features by the participants. Feedback from the participants confirmed this and brought up some interesting issues. Participants reported that they were hesitant about taking food pictures in public since it

attracted undue attention. They also pointed out that Indians mostly eat with their hands/fingers and take several food servings, which make it difficult to take pictures of subsequent servings. These social issues need to be taken into account in the next design. There were no usability concerns with regard to Dia-Graph. The co-PI found the software easy-to-use, and the interface to be efficient.

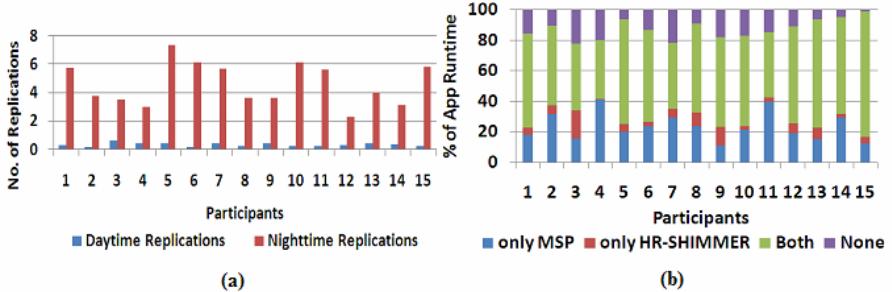


Fig. 11. (a) Average replications per day and (b) Average percentage sensing coverage

4.2.3 System Performance

We faced minor technical problems during the deployment. Initially participants reported a number of problems with the HR-SHIMMER like connection problems with the phone and low battery life. As a result, all the HR-SHIMMERs were recalled, tested and treated for moisture protection at the end of two weeks.

Initially, a few participants complained about too many synchronization attempts occurring during the daytime resulting in interrupting system usage. This was due to the fact that the system was set up to attempt a daytime synchronization if it failed the previous night. When we changed the condition for a daytime synchronization to failure on 3 consecutive nights, participants experienced far less problems. Overall, we found that the number of daytime synchronizations was less than 1 on average (Fig. 11(a)) and the total time taken by synchronization over the study was 10 hours on average per participant. Synchronization took an average of 26 minutes.

Since some of the participants reported that their sensors often ran out of charge, we analyzed the data logged on the phones to estimate daily sensing coverage. The results of our analysis are shown in Fig. 11(b). Participants were covered by both or either sensor for about 87% of the application runtime, on average. We find this to be a high percentage, given that sensors were generally disabled during the synchronization process. An important point to note here is that participant 15, who regularly charged the sensors as instructed, had a sensing coverage of more than 99%. This shows that the system worked as expected when the protocol was strictly followed. Finally we found the MSP to last longer than the HR-SHIMMER.

4.2.4 Usefulness

Feedbacks from the participants like “*Makes one self conscious; acts as a constant reminder,*” “*Really got conscious of calorie contents of various items,*” and “*After using these gadgets realized that I end up eating more than required almost daily for I always thought that I was a poor eater,*” show that the system was helpful in various

ways. While for some, it served as a constant reminder, for others, it helped improve their understanding of their diet and activity habits. In fact, 13 participants reported to have gained a better understanding of their diet and activity habits after using the system. All participants claimed that it helped improve their diet while all but one said that it helped increase their physical activity.

We also evaluated the usefulness of each of the system features. Participant feedback revealed that they found all the application features useful to varying degrees. The EE and CI bars seemed to be the most popular followed by heart rate information and trends display. Participants reported viewing the MyDay screen and the trends screen frequently, ranging from 2-5 times per day to more than 10 times per day. Majority viewed the MyDay screen greater than 10 times a day and the trends screen 2-5 times a day. One participant commented "*I keep checking the calorie count as frequently as checking the watch for time.*" Thus, overall, participants seemed to find the system useful and were willing to continue use of the system over a long-term. 10 participants felt that the change produced by the system will continue even after the study. Almost all except one said they would recommend this type of system to their close friends who are in need of better life style management.

4.3 Lessons Learned

This study brought to light several insights. First and foremost, such long-term-wear systems should be unobtrusive and easy to wear. Moreover they should involve as few components as possible (for easy carrying and charging). We realize that though at present, wearing multiple sensors and the chest belt is cumbersome for everyday use, we believe that in future these sensors would become smaller, energy efficient, more wearable and integrated into a single device (e.g. embedded in clothing, watch or a phone). We also learned that such systems should not attract undue attention to the wearer in a public setting (as with the blinking lights on the MSP and the need to photograph the food in public). To improve robustness, the system should gracefully handle sensor disconnection (e.g when sensors are separated from the phone). Caching the data on the sensors and transferring the data later when the connection to the phone is reestablished is a possibility. We also found that calorie intake and energy expenditure information, and caloric values of individual food items are key to influencing behavior change in individuals suffering from metabolic syndrome based chronic diseases. While participants were excited about long-term use of the device, when asked if they might get bored of the device in the long run, 6 said that they weren't sure while 4 said that it was likely that they would. Hence, it's critical that user-fatigue be considered while designing such a device in order to support long term use. Finally, access to granular user activity and diet information help physicians provide personalized and better quality of care to individuals.

5 Conclusion

Diabetes is emerging as a major public health concern. Empowering the user to lead a healthy lifestyle by increasing physical activity and moderating eating habits plays a key role in self-management of this syndrome. Lifestyle is highly diverse and dependent on cultural factors. To be effective, a pervasive lifestyle management system has

to be comprehensive as well as customized to personal preferences. There have been studies in developed countries regarding efficacy of self-management intervention in lifestyle modification using pervasive healthcare with mixed results. India is on the verge of being the diabetes capital of the world, which calls for a comprehensive study in the Indian context.

Though recent clinical studies have indicated that increased physical activity and diet modification can effectively retard diabetes progression, currently there is no reliable system available to objectively monitor and manage their physical activity, calories spent and food consumed on an ongoing basis. Hence, to overcome this deficiency and empower the patients with a reliable system that provides actionable information, Jog Falls was developed and evaluated in a user study. Objectives of the study were to evaluate the reliability, acceptability, usability and usefulness of the system and to improve the system based on the observations, results and feedback.

Overall, the system was accepted by the users, was used extensively and emerged as an easy to use system for lifestyle management. Self-awareness of caloric values of food and ease of calorie intake/expenditure logging were well appreciated by the users. Statistically significant weight reduction was observed in users and there was very strong correlation between hours of use of the system and weight reduction. The research evaluation revealed that simplification of food data entry and automation of calorie intake would improve the usability of the system. Additionally, reduction in form factor, number of devices, simple charging mechanisms and longer battery life would further simplify the use of the system.

Acknowledgement

The authors would like to acknowledge the National Institute of Nutrition in Hyderabad, India for providing access to the Indian Food Database. We would also like to acknowledge Deniz Arik from TU Delft for providing valuable input on the User Interface of Jog Falls based on his user studies at Manipal University.

References

1. Joshi, S.R., Parikh, R.M.: India – Diabetes Capital of the World: Now heading towards Hypertension. *Journal of the Association of Physicians of India*, 323–324 (2007)
2. Yamax, http://www.yamaxx.com/digi/cw_200_e.html
3. Nike+iPoD, <http://www.apple.com/ipod/nike/run.html>
4. Polar Actiwatch,
http://www.polarusa.com/us-en/products/fitness_crosstraining/AW200
5. Fitbit, <http://www.fitbit.com>
6. BodyBugg, <http://www.bodybugg.com/>
7. Polar Heart Rate Monitor,
http://www.polarusa.com/us-en/products/fitness_crosstraining/F4
8. Consolvo, S., Klasnja, P., McDonald, D.W., Avrahami, D., Froehlich, J., LeGrand, L., Libby, R., Mosher, K., Landay, J.A.: Flowers or a Robot Army? Encouraging Awareness & Activity with Personal, Mobile Display. In: *Proceedings of the 10th International conference on Ubiquitous computing*, pp. 54–63 (2008)

9. Choudhury, T., Consolvo, S., Harrison, B., Hightower, J., LaMarca, A., LeGrand, L., Rahimi, A., Rea, A., Bordello, G., Hemingway, B., Klasnja, P., Koscher, K., Landay, J.A., Lester, J., Wyatt, D., Haehnel, D.: The Mobile Sensing Platform: An Embedded Activity Recognition System. *IEEE Pervasive Computing* 7(2), 32–41 (2008)
10. Consolvo, S., Everitt, K., Smith, I., Landay, J.A.: Design requirements for technologies that encourage physical activity. In: 24th international conference on Human factors in computing systems, pp. 457–466 (2006)
11. Maitland, J., Sherwood, S., Barkhuus, L., Anderson, I., Hall, M., Brown, B., Chalmers, M., Muller, H.: Increasing the awareness of daily activity levels with pervasive computing. In: *Pervasive Health*, pp. 1–9 (2006)
12. Lin, J.J., Mamykina, L., Lindtner, S., Delajoux, G.: Fish'n Steps: encouraging physical activity with an interactive computer game. In: Dourish, P., Friday, A. (eds.) *UbiComp 2006*. LNCS, vol. 4206, pp. 261–278. Springer, Heidelberg (2006)
13. Nutrition Vista, <http://nutritionvista.com>
14. Body Media, <http://bodymedia.com>
15. FitDay, <http://www.fitday.com>
16. Myfoodphone, <http://www.myfoodphone.com>
17. Mamykina, L., Mynatt, E., Davidson, P., Greenblatt, D.: MAHI: investigation of social scaffolding for reflective thinking in diabetes management. In: 26th international conference on Human factors in computing systems, pp. 477–486 (2008)
18. Weight Watchers, <http://www.weightwatchers.com/index.aspx>
19. Bao, L., Intille, S.S.: Activity Recognition from User-Annotated Acceleration Data Export. In: *Pervasive Computing*, pp. 1–17 (2004)
20. Brage, S., Brage, N.: Branched equation modeling of simultaneous accelerometry and heart rate monitoring improves estimate of directly measured physical activity energy expenditure. *J. Appl. Physiol.* 96, 343–351 (2004)
21. Locke, E.A., Latham, G.P.: A theory of goal setting and task performance. Prentice Hall, Englewood Cliffs (1990)
22. MSP Platform description, <http://seattle.intel-research.net/MSP>
23. SHIMMER, http://shimmer-research.com/wordpress/?page_id=20
24. Treuth, M.S., Adolph, A.L.: Energy expenditure in children predicted from heart rate and activity calibrated against respiration calorimetry. *American Physiological Society* 275, E12–E18 (1998)
25. Spurr, G.B., Goldberg, G.R.: Energy expenditure from minute-by-minute heart-rate recording: comparison with indirect calorimetry. *American Journal of Clinical Nutrition* 48, 552–559 (1988)
26. Ainsworth, B.E., Haskell, W.L., Whitt, M.C.: Compendium of physical activities: An update of activity codes and MET intensities. *Med. Sci. Sports Exerc.* 32, S498–S516 (2000)
27. Andre, D.: The Development of the SenseWear® armband, a Revolutionary Energy Assessment Device to Assess Physical Activity and Lifestyle. BodyMedia Inc. (2006)
28. Olgun, D., Pentland, A.: Human activity recognition: Accuracy across common locations for wearable sensors. In: Proceedings of the IEEE 10th International Symposium on Wearable Computing (Student Colloquium Proceedings) (2006)
29. Consolvo, S., McDonald, D.W., Landay, J.A.: Theory-Driven Design Strategies for Technologies that Support Behavior Change in Everyday Life. In: 27th international conference on Human factors in computing systems, Boston, MA, pp. 405–414 (2009)
30. Fogg, B.J.: Persuasive Technology: using computers to change what we think and do. Morgan Kaufmann Publishers, Boston (2003)
31. Intille, S.S.: A New Research Challenge: Persuasive Technology to Motivate Healthy Aging. *IEEE Transactions on Information Technology in Biomedicine*, 235–237 (2004)

EyeCatcher: A Digital Camera for Capturing a Variety of Natural Looking Facial Expressions in Daily Snapshots

Koji Tsukada and Maho Oki

Ochanomizu University, 2-1-1, Otsuka, Bunkyo-ku, Tokyo, 112-8610, Japan

{tsuka,okimaho}@acm.org

<http://mobiquitous.com/>

Abstract. This paper proposes a novel interactive technique, the EyeCatcher, which helps photographers capture a variety of natural looking facial expressions of their subjects, by keeping the eyes of the subjects focused on the camera without the stress usually associated with being photographed. We develop a prototype system and verify the effectiveness through evaluation and discussion.

1 Introduction

As digital cameras have become increasingly popular in recent years, people have come to take more pictures in their daily lives. In particular, people often take snapshots of their families, friends or pets. However, many people experience difficulties in capturing the natural facial expressions of their subjects for several reasons: many subjects become stressed when facing a camera, while other “camera-wise” subjects –those accustomed to being photographed– often make stage faces. Moreover, it is often quite difficult to take pictures of children, since they often look away from the camera.

This paper proposes a novel interactive technique, the EyeCatcher, to help photographers capture a variety of natural looking facial expressions by keeping the eyes of subjects focused on the camera without the stress of photography (Fig. ).

2 EyeCatcher

The main concepts of the EyeCatcher are as follows:

1. Keeping the eyes of the subjects focused on the camera
2. Reducing the stress associated with being photographed
3. Extending existing digital cameras

First, the EyeCatcher can help to keep the eyes of the subject focused on the camera. Some people who are not comfortable with being photographed often

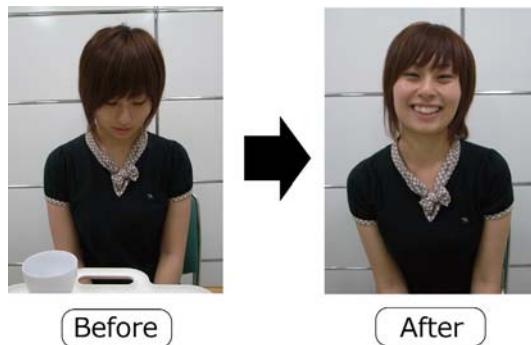


Fig. 1. The goal of the EyeCatcher is to help photographers capture a variety of more natural looking facial expressions of subjects by keeping the eyes of the subjects focused toward the camera without stress

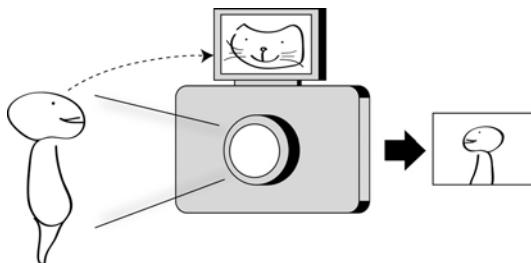


Fig. 2. The basic concept of the EyeCatcher is to keep the eyes of the subjects focused toward the camera, while turning their attention to the content shown in a small display attached above the lens on the front of the camera

turn their eyes away when they are faced with a camera. Moreover, since children tend to move around restlessly, photographers often experience difficulty in keeping their eyes towards the camera.

Second, the EyeCatcher can reduce “camera-stress” on subjects. When photographers take pictures of subjects, many subjects become stressed since their attention is centered on the camera; other “camera-wise” subjects –those accustomed to being photographed– often create stage faces, and their faces look almost the same in every picture.

To solve these problems, we attach a small display to the front of the camera. By presenting images or videos (e.g., friends, pets, or animation characters) on the display, we can (1) keep the eyes of the subjects focused toward the camera, and (2) turn their attention to content shown in the display and away from the stress of being photographed(Fig. ②).

Third, we can extend the versatility and practicability of existing digital cameras. Since many photographers have their favorite cameras, it is likely they



Fig. 3. Hot shoe connector of a high-end compact digital camera (Ricoh GR Digital2). There are 5 signal terminals at center, and ground terminals on each side.

would like to apply this innovation to their current cameras rather than use completely new ones. For this reason, we designed the system so that the small display can be attached to existing digital cameras using a “hot shoe connector”. The hot shoe connector is an extension connector mainly used for strobes by experienced photographers on many digital cameras; both single lens reflex cameras and high-end compact cameras (Fig. 3).

The function of the hot shoe connector is to connect the camera with an external device both “physically” and “electrically”. For example, when a strobe is attached to the camera via the hot shoe connector, the camera can transmit many commands to the strobe via electrical signals, thereby controlling apertures, shutter speeds, zooms, shutter button status, and so on.

Using a hot shoe connector to attach our novel device, we can not only stably fit the device on the camera, but also detect input signals from the camera (e.g., shutter button status) and use them to control the device. Moreover, we can avoid parallax problems using the hot shoe connector since it is usually located directly above the lens, as will be explained in greater detail in the “discussion” section.

3 Implementation

In this section, we explain the implementation of the EyeCatcher prototype. First, we selected a high-end compact digital camera (Ricoh GR Digital2) for attachment of the EyeCatcher. The GR Digital2 is famous for its picture quality, and is used extensively as the camera of choice by professional and semiprofessional photographers. Fig. 4 shows an image of the prototype.

The prototype system consists of three main components: (1) a presentation component on the front, (2) a selection component on the back, and (3) a control component located between the two.

The presentation component consists of an organic EL display, 4D Systems uOLED-160-G1 (Fig. 5 bottom left). The uOLED-160-G1 is a full-color organic



Fig. 4. The prototype EyeCatcher

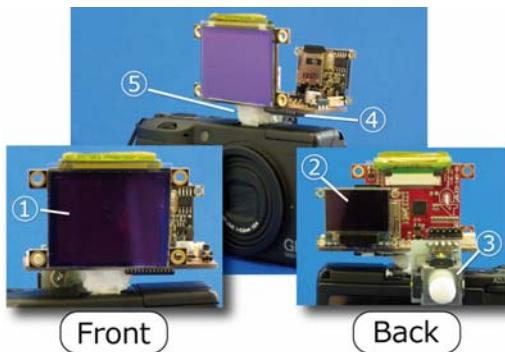


Fig. 5. The system architecture of the prototype. 1. organic EL display (uOLED-160-G1), 2. organic EL display (uOLED-96-Prop), 3. joystick, 4. micro controller(PIC18F2550), 5. hot shoe connector.

EL display. The resolution and size are 160 x 120 pixels and 32mm x 40mm, respectively. It has much higher visibility than ordinary LCDs; i.e, it is much brighter and has a wider angle of view (about 180 degrees). Dots, lines, shapes, text, and image or video content can easily be displayed by sending several bytes of commands via a UART communication link. Moreover, it has a micro SD slot as an external memory, so we can easily update content using an everyday personal computer. We use the uOLED-160-G1 for displaying various content to the subjects.

The selection component consists of an organic EL display, 4D Systems uOLED-96-Prop, and a joystick, CTS 252A103B60NA (Fig. 5 bottom right). The uOLED-96-Prop is a full-color organic EL display, with specifications almost the same as those of the uOLED-160-G1, apart from resolution and size (96 x 64 pixels and 23mm x 25.7mm). The joystick detects the movement of the stick with 2 variable resistors, and outputs 2 analog signals. It also works as a push button by pressing the stick. We use the joystick and the uOLED-96-Prop

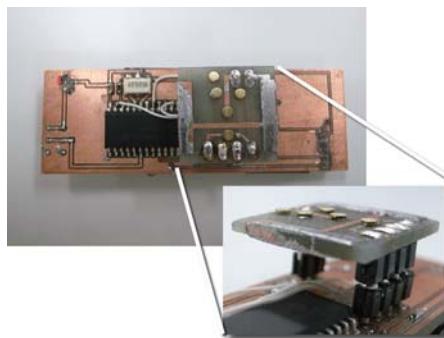


Fig. 6. The control board and hot shoe connector for the prototype

EL display for selecting content for visual feedback, as will be explained in detail in the following section.

The control component consists of a micro controller (Microchip Technology PIC18F2550), an original hot shoe connector, and peripheral circuits (Fig. 6). The size of the board is about 22mm x 65mm¹. The hot shoe connector is designed for the hot shoe socket of the GR Digital2 shown in Fig. 3². As mentioned above, the hot shoe connector works (1) to physically attach the EyeCatcher device to the camera and (2) to electrically connect the signal lines between the two components. Thus, the EyeCatcher can detect several camera operations (e.g., the pressing of the shutter button only halfway) by analyzing signals from the hot shoe connector. The control component is connected to the presentation and selection components via pin headers, and controls these devices via the micro controller. We use a lithium ion battery (3.7V) as a power supply³.

Finally, we developed the outer package of the prototype using ABS plastics. Since we have integrated all components into the outer package, the EyeCatcher device can be easily be connected or removed like an external strobe.

3.1 Content

In this section, we explain the content shown in the EyeCatcher. We define the conditions for selecting content as follows:

¹ This size is smaller than the upper surface of the GR Digital2.

² The layout of the electrical contacts of the hot shoe connectors vary between manufacturers. Although this prototype works only with Ricoh cameras, we believe it would not be difficult to support cameras from other manufacturers by designing corresponding connectors.

³ We initially planned to supply power from the hot shoe connector, since there was a contact that supplies about 3V. However, since this voltage was not sufficient to operate the organic EL displays (which need 3.6V), we passed over this idea in the current prototype.

1. Content for attracting the attention of subjects
2. Content familiar with subjects
3. Content for producing various expressions or poses

The first point is to attract the attention of the subjects at a glance. For example, displaying images of human faces is better suited to keeping the subjects' attention [1]; whereas animation is usually more attractive than still images (Fig. 7 left top). Moreover, since the display size of the current prototype is relatively small, a simple composition may be more desirable than a complex one.

Secondly, EyeCatcher uses content that is familiar to the subjects as an attempt to invoke the most reaction. For example, the system uses pictures of friends or associates rather than those of complete strangers (Fig. 7 top right). The system also uses pictures of actors, artists or animation characters which are well-known to many people.

The third point is that some contents may help people produce various expressions or poses. For example, silhouettes of poses may trigger unique poses (Fig. 7 bottom right), while face icons may help subjects produce similar faces (Fig. 7 bottom left).

Figure 7 shows examples of the content which meet the above conditions.



Fig. 7. Examples of content for use in the EyeCatcher

Next, we explain how users select the content to present to subjects with EyeCatcher. We designed a virtual matrix menu suited for control with the joystick. Users can directly select the menu by moving the stick in any of 8 directions. The menu consists of 2 hierarchies: there are 8 category folders and each of them has 8 options (Fig. 8).

The procedures for selecting and presenting content are as follows:

1. First, the photographer browses the categories for content by moving the joystick in any of the 8 directions. The category title and a typical image of the selected category are shown in the back display.

2. When the joystick is kept in the same direction for a few seconds⁴, the category folder opens.
3. Next, the photographer can browse the options within the selected folder by moving the joystick again. The title and image of the selected option are again shown in the back display. The user can return to the category menu by pressing the joystick.
4. After selecting the content for display, the photographer points the camera at the subject, and presses the shutter button halfway. Since the status of the shutter button is automatically detected by the system, the selected image is shown in the front display (Fig. 9)⁵. Moreover, since the selected content is shown to the subjects just before the picture is taken, the photographer can easily capture the reactions of the subjects.
5. A few seconds⁶ after the shutter button is pressed, the content shown in the front display is cleared.

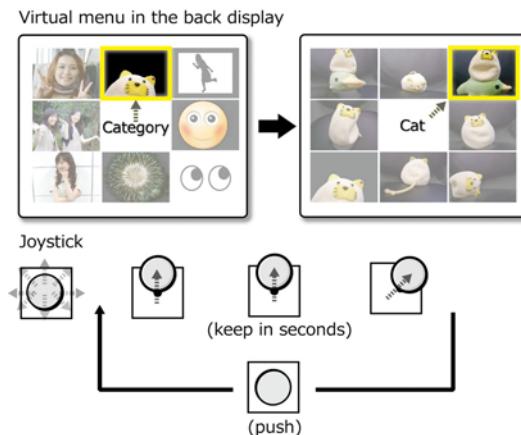


Fig. 8. Procedure for selecting content. Users can browse contents in the matrix menu by moving the joystick in 8 directions. They can select a content option by keeping the stick in the same direction for a few seconds. Information regarding the current category or content is shown in the back display.

We have designed these control procedures to be easily used with the thumb when the photographer is holding the camera in his or her hands. Moreover, when the photographer learns the menu structure, he or she may select content without looking at the back display. Therefore, this method may also be useful for single lens reflex cameras, which are mostly used with optical finders⁷.

⁴ 1 second in the current prototype.

⁵ Before this step, content is only shown in the back display; alternatively, several animated lines, like a screen saver, are shown in the front display.

⁶ 2 seconds in the current prototype.

⁷ In future implementations, we intend to include feedback functions using clicks or vibrations.

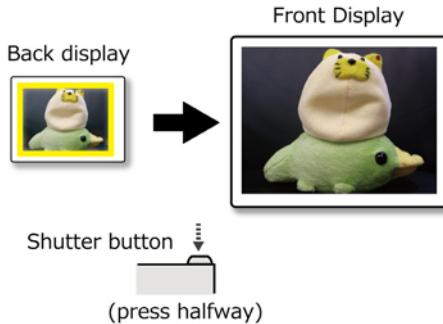


Fig. 9. Procedure for displaying content. After previewing a selected content option in the back display, the user can show it in the front display by pressing the shutter button halfway.

4 Evaluation

We evaluated the effectiveness of the EyeCatcher prototype with regard to two aspects: (1) “How the subjects feel about the EyeCatcher?” and (2) “How people feel about the pictures captured using the EyeCatcher?”. First, we took pictures of subjects while showing content using the EyeCatcher, and then obtained subjective feedback from the subjects via a questionnaire. Next, we conducted another questionnaire survey to examine impressions of the captured pictures. We define the subjects of the first evaluation as “subjects” and the subjects of the second evaluation as “respondents” to avoid confusion.

4.1 Photography Evaluation

Method. We selected eight test subjects (7 female and 1 male) from among the members of our laboratory who had never used the EyeCatcher before. Their ages ranged between 22 and 52. All participants use of digital cameras (including camera-equipped cell phones) on a daily basis.

The experimenter took each subject into a room, seated her/him on a chair, and seated himself across a table from the subject. The distance between the subject and the experimenter was about 1m. This distance was decided in consideration of the characteristics of the digital camera used in the current prototype (GR Digital2). As the GR Digital2 is equipped with a wide-angle fixed-focus lens (28mm), subjects in the pictures taken from farther than 1m appeared too small for our purposes.

The procedure for this evaluation was as follows. First, the experimenter took pictures of each subject using the GR Digital2 without the EyeCatcher. Next, the experimenter took pictures using the EyeCatcher while showing eight content options in a random sequence. The content options used in this evaluation are shown in Fig. 10.⁸

⁸ Content 4 was a simple animation, and all others were still images.



Fig. 10. The contents used in the evaluation. 1. Food (cake), 2. Face icon (surprised), 3. Japanese actor (Kenichi MATSUYAMA), 4. Animation character, 5. Associate (professor), 6. Pose (hands on the waist), 7. Japanese entertainer (Harumi EDO), 8. ID photo (man wearing a suit).

These content options were selected based on the conditions mentioned in the “Content” section. That is, (1) all content should have a clear composition for attracting the attention of subjects at a glance, (2) some of them should be familiar to the subjects, and (3) others should help the subjects produce various expressions or poses. We selected content 1, 3 and 4 based on the condition (2), and content 2, 6 and 8 on condition (3). Content 5 and 7 were selected as fulfilling both condition (2) and (3). Since the subjects were mostly young women, we selected “cake” as a food, and a “young male actor” as an actor taking into account their preferences.

The experimenter did not engage in any verbal communication with the subjects to reduce variables other than the effect of the EyeCatcher. To begin, the experimenter told each subject “Please act as you usually would, and don’t be nervous about the experiment.”. After starting the evaluation, the experimenter did not speak to the subjects apart from briefly replying to the subjects’ questions. When the evaluation was finished, the experimenter obtained subjective feedback from the subjects both by questionnaires and discussion.

Figure 11 shows pictures taken using the GR Digital2 without the EyeCatcher.⁹

Results. First, we attempted to characterize the subjects by asking them two questions: (1) “How comfortable are you with being photographed? (1: very uncomfortable - 5: very comfortable)”, (2) “How emotionally expressive are you

⁹ Pictures shown in this paper are cropped for greater visibility.



Fig. 11. Pictures captured without the EyeCatcher

when photographed? (1: very unexpressive - 5: very expressive)¹⁰. Figure 12 shows the distribution of the subjects in terms of their response to these two questions. The subjects were found to be divided into three main groups: (1) subjects E and G were “comfortable with being photographed” and “expressive”, (2) subjects B, F and H were “uncomfortable with being photographed” and “inexpressive”, and (3) subjects A, C and D fell between those in groups (1) and (2). We will refer to subjects E and G as “camera-wise subjects” and subjects B, F, and H as “camera-shy subjects,” and discuss the results based on these terms in the “Consideration” section.

Next, we explain the results of the subjective questionnaire. We set four questions and scored the answers on a scale of 1 to 5 as follows: (1) “Was your focus on the camera reduced? (1: not reduced at all – 5: drastically reduced)”, (2) “Was the photography process pleasant? (1: very unpleasant – 5: very pleasant)”, (3) “Were your faces captured differently? (1: not different at all – 5: completely different)”, (4) “Would you want to always use the EyeCatcher in the future? (1: would never want to – 5: extremely want to)”.

Figure 13 shows the results of this questionnaire. For question (1), 7 of 8 subjects answered that their attention to the camera was reduced ($\text{avg}=3.75$, $S.D.=1.39$). For question (2), 7 of 8 subjects said the photography process was pleasant ($\text{avg}=4.25$, $s.d.=1.03$). For question (3), 7 of 8 subjects felt their faces were captured differently when the EyeCatcher was used ($\text{avg}=4.13$, $s.d.=0.99$). For question (4), 7 of 8 subjects answered that they would want to always use

¹⁰ All questions and answers presented in this paper are translations from the original Japanese.

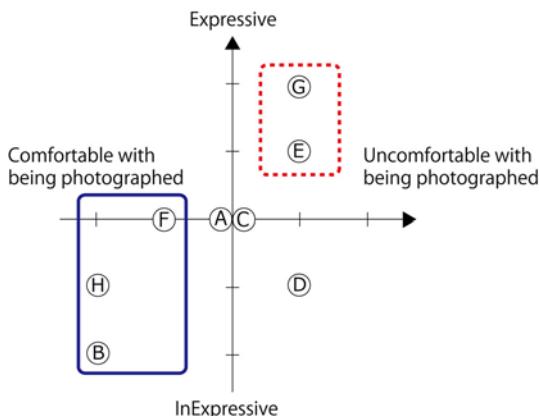


Fig. 12. Character distribution of the subjects

the EyeCatcher in the future ($\text{avg}=4.25$, $\text{s.d.}=0.71$). In addition, 2 of the 3 camera-shy subjects answered 4 or higher for all the questions. The self-reported characteristics of the subjects as assessed initially appeared not to affect their responses to the questions that they were asked after the experiment. Thus, we argue that the EyeCatcher was effective from the subjects' perspective.

4.2 Impression Evaluation

Next, we evaluated people's impression on the resulting images to verify the effect of the EyeCatcher on the captured pictures.

Method. First, we showed respondents (1) pictures taken without the EyeCatcher ("regular pictures") and (2) pictures taken with the EyeCatcher ("EyeCatcher pictures") at the same time, and obtained feedback regarding their impression of the EyeCatcher pictures compared to the regular pictures. We prepared 64 EyeCatcher pictures (8 subjects x 8 content options) and 8 regular pictures (8 subjects). The sequences in which the pictures were shown were changed randomly. The respondents were not shown the content used in the photography evaluation. We selected 9 respondents with ages ranging between 21 and 37 who were not included as subjects in the earlier evaluation.

Results. We asked 2 questions and scored answers on a scale of 1 to 5 as follows: (1) "Do you notice any difference between the EyeCatcher pictures and regular pictures? (1: not different at all – 5: completely different)", (2) "Do you feel the EyeCatcher pictures are less strained compared to the regular pictures? (1: very strained – 5: very unstrained)". We calculated average scores and standard deviation for all pictures.

The results for question (1) are shown in Fig. 14. The horizontal axis shows EyeCatcher pictures, and the vertical axis shows average score. First, almost all

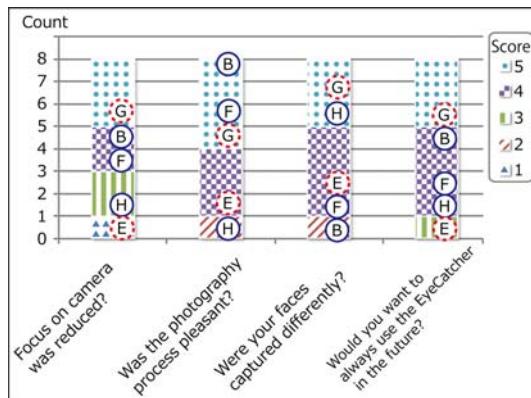


Fig. 13. Results of the subjective evaluation

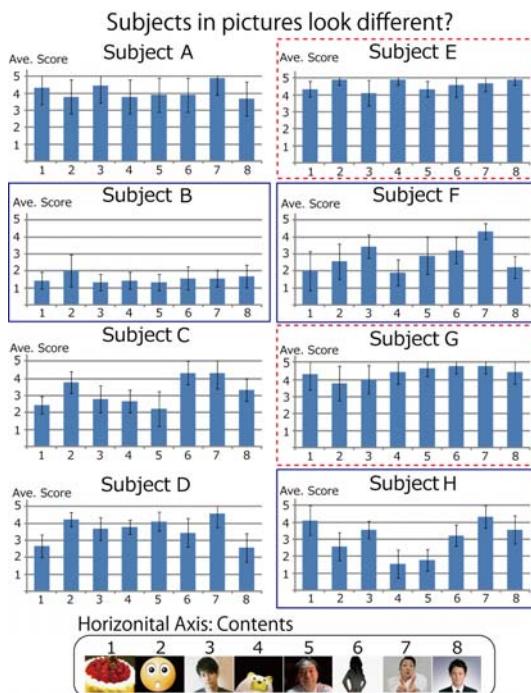


Fig. 14. Do you notice any difference between the EyeCatcher pictures and regular pictures?

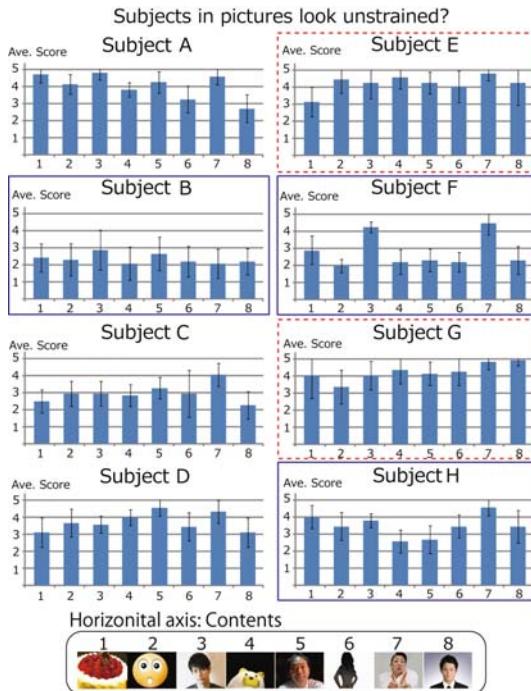


Fig. 15. Do you feel the EyeCatcher pictures are less strained compared to the regular pictures?

of the pictures of subjects A, E, and G obtained high scores (above 4.0). Most of the respondents felt the EyeCatcher pictures were “5: completely different” or “4: rather different” to the regular pictures. Second, scores for the pictures of subjects C, D, F and H were varied, particularly, those for subjects F and H for whom some pictures were regarded as “different”(above 4.0) whereas others were regarded as the “same”(under 2.0).

Next, The result of the question (2) is shown in Fig 15. First, 7 of 8 pictures of the subjects E and G obtained high score (above 4.0). Most respondents felt their EyeCatcher pictures “5: very unstrained” or “4: unstrained”. Second, almost all pictures of subjects A and D obtained reasonably high scores above 3.5. Many respondents felt their pictures to be “4: unstrained”. Third, scores for the pictures of subjects C, F, and H were varied. There were both “unstrained” pictures (above 4.0) and “little strained” pictures (under 3.0) for each subject. In particular, some pictures of subject F were felt to be “strained” (under 2.0).

Meanwhile, pictures of subject B obtained low scores for both questions. We couldn't observe any effectiveness of the EyeCatcher for subject B.

We have also asked the respondents about their relationships to the subjects. We found that 4 knew all subjects, 2 knew some of them, and 3 respondents hardly knew them. Although we have not fully analyzed the results by group, no significant differences appear to exist.



Fig. 16. Examples of high-scoring pictures (content 7, Japanese entertainer)

4.3 Consideration

In this section, we consider the results of the evaluations. The EyeCatcher pictures that obtained (1) higher and (2) lower scores in the impression evaluation are as follows: (1) pictures of content 7 (Japanese entertainer, Harumi EDO) shown in Fig. 16 and (2) pictures of content 8 (suited man) shown in Fig. 17.

First, we consider the result of the impression evaluation based on the self-reported characteristics of the subjects shown in Fig. 12. For camera-wise subjects (subjects E and G), most pictures obtained high scores (above 4) both in terms of “difference” and “unstrained”. This result indicates that the EyeCatcher helps a photographer capture various natural expressions of camera-wise subjects. In the regular pictures (Fig. 11), camera-ready subjects made stage faces and paid much attention to the camera. In contrast, they reacted to most of the content shown using the EyeCatcher since they were very emotional. For this reason, we could capture various expressions and poses for these subjects. In response to content 7 (Fig. 16), for example, subjects E and G struck a similar pose to that of the entertainer¹¹. Meanwhile, we selected content 8 (Fig. 17) for taking pictures suited to use on ID cards. However, the attention of most subjects was focused on “who is this person?”, and they became confused. For this reason, the scores of the impression evaluation were lower¹². Although subjects E and G also felt the same impression, they reacted dynamically: subject E bent forward to the camera and made a confused face; subject G began laughing helplessly.

¹¹ Although subjects A, C and D also struck a similar pose, subjects E and G did not only strike a similar pose, but also made a similar face.

¹² Only subject D understood our intention and stood erect.



Fig. 17. Examples of low-scoring pictures (content 8, suited man)

For these reasons, even the pictures taken using content 8 obtained high scores. Although we didn't expect these reactions, we thought them to demonstrate the interesting effects possible using the EyeCatcher.

Next, we discuss the camera-shy subjects (subjects B, F and H). The scores for the pictures of subjects F and H were spread. For example, for the “difference” question, pictures taken using content 4 were felt to be “hardly different” from the regular pictures (below 2.0); whereas the pictures taken using content 3, 7 (both subjects F and H) and 1 (subject H) were felt to be “different” from the regular pictures (above 3.5). For the “unstrain” question, pictures taken using content 4 and 5 were felt to be a “little strained” compared to the regular pictures (below 2.7); whereas pictures taken using content 7 (both subjects F and H), 3 (subject F) and 1 (subject H) were felt to be “unstrained” (above score 4.0). Thus, we could capture natural and unstrained faces when certain content (3 and 7 for subject F; 1 and 7 for subject H) were shown (Fig. 16). In summary, (1) the faces of subjects F and H were strained in the regular pictures (Fig. 11) since they were uncomfortable with being photographed and (2) they did not react to all the content since they were unexpressive. However, when the preferred content (e.g., cakes, actors, and entertainers) were shown, the EyeCatcher could help the photographer capture natural smiles even on camera-shy subjects.

Next, we consider subject B. As mentioned above, the EyeCatcher was not effective for subject B. We think that there were three reasons for this: (1) subject B was the most unexpressive and the most uncomfortable at being photographed; (2) the content options were ill-suited to subject B as his sex and age (male, 52) were different from those of the other subjects; and (3) subject B did not change his face intentionally since he misunderstood the instruction to “please

act as you usually would” as asking him gnot to change his expression from his usual look”. However, in the subjective feedback, subject B did respond by saying that “his focus on the camera was reduced” and “photography process was pleasant”. Thus, we believe the EyeCatcher can capture natural and unstrained expressions even for subject B in future by providing more suitable content and communicating verbally with the subject.

Thus, the EyeCatcher can help a photographer (1) capture various natural expressions and poses of camera-wise subjects and (2) capture the natural smiles of camera-shy subjects by showing preferred content.

5 Discussion

In this section, we discuss the basic performance of the EyeCatcher and communication during photography.

5.1 Visibility of Display

The performance of the EyeCatcher in the field is somewhat influenced by the visibility of the front display. Therefore, we tested the EyeCatcher under several sets of conditions and verified the visibility of the display. We selected the 4 content categories (characters, friends, face icons, and poses) shown in Fig 7 and examined the distance at which a subject could easily recognize the content. The subject had normal eyesight. Results showed that the subject could easily recognize all contents from 2 m in a room lit with fluorescent lamps, 1.5 m outside on a sunny day but without direct light, and 1 m outside on a sunny day with direct light.

Since most snapshots are usually taken from 1-3 m, the visibility of the current prototype appears to be practicable.

5.2 Correspondence of Eyes

Generally, in systems equipped with both cameras and displays (e.g., TV conference systems), the focus of the eyes often becomes a problem since the location of each device is different [2]. As it was thought that the EyeCatcher may experience a similar problem, we discuss this topic here.

Minami [2] reported the detection limit of correspondence of eyes is about 2 degree and the allowable limit is about 9 degree. In the EyeCatcher, since the distance between the lens and the display is about 6 cm, the allowable limit is crossed when a subject comes closer than 38 cm. However, we think that this will not become a significant problem since few photographers take photographs of their subjects at such close range. For example, when the distance between the EyeCatcher and a subject is 1m, the parallax is about 3.4 degrees, which is much smaller than the allowable limit. Moreover, there were almost no pictures taken during the evaluation in which we can observe problems associated with the focus of the eyes.

5.3 Communication in Photography

Finally, we discuss communication during photography. When taking snapshots of people, communication between the photographer and subject is quite important since the photographer cannot control the subject's face directly. For example, professional photographers do not only require various photographic techniques, but also communication skills to reduce stress or indicate the pose he feels will be most attractive.

The goal of the EyeCatcher is to capture various natural expressions of the subject, and it supports communication between the photographer and subject by creating a new communication channel via the visual content. For example, it's usually quite difficult for photographers to pose subjects similar to that seen in content 6 or 7 in Fig.10 with only verbal instruction. The EyeCatcher offers useful solutions for such situations. Meanwhile, from the comments received during the photography evaluation, some subjects would like to receive verbal instructions such as "Please mimic it!". Thus, verbal communication is also important for photography using the EyeCatcher.

6 Related Work

From the results of PC-based experiments, CheeseCam³ reported unconscious reactions of subjects when watching face icons. Based on this research, Samsung released a digital camera (DualView TL225) with a front display⁴. Similarly, Howdy⁵ is a unique digital camera that looks like a photo frame. It can capture pictures of the subject and the photographer at the same time using small cameras attached on both sides of the frame. Since they can look at each others faces, the photographer can capture less strained images of the subject. These approaches share the same goal as that of the EyeCatcher, reducing the degree of attention that the subject pays to the camera and thereby capturing the subjects natural expression.

The uniqueness of the EyeCatcher is that (1) it is easily attached to existing digital cameras and (2) it allows subjects to produce various expressions by the easy changing of displayed content.

There are several digital cameras that have integrated sensors. ContextCam⁶ proposed a context-aware video camera that provides time, location, person presence and event information. Likewise, WillCam⁷ helps the photographer capture various information, such as location, temperature, ambient noise, and photographer's facial expression, in addition to the photo itself. Capturing the Invisible⁸ designed real-time visual effects for digital cameras using simulated sensor data. The EyeCatcher focuses on photography process itself, and intends to capture the various natural expressions observed in our daily lives.

7 Conclusion

This paper proposed a novel interactive technique, the EyeCatcher, which helps photographers capture various natural expressions on their subjects, by keeping

eyes of subjects focused toward the cameras without the stress often associated with being photographed. We developed a prototype system that can be attached using the hot shoe connector found on existing digital cameras. Moreover, we verified the effectiveness of the EyeCatcher through evaluation and discussion. Our study population was small in scale, unbalanced in composition and consisted solely of members of our laboratory, so that they might have been unduly supportive of the system. Nevertheless, our findings offer positive, if only preliminary, data regarding the potential value of the EyeCatcher system.

References

1. Kanwisher, N.: What's in a face? *Science* 311
2. Minami, T.: Art of video telephone. *IEICE Transactions* 56(11), 1485–1490 (1973) (in Japanese)
3. Lee, B., Lee, W.: Cheese cam: unconscious interaction between humans and a digital camera. In: Extended abstracts on CHI 2009, pp. 4285–4290. ACM Press, New York (2009)
4. Samsung dualview tl225 (2009),
http://www.samsung.com/us/consumer/photography/digital-cameras/compact/EC-TL225ZBPOUS/index.idx?pagetype=prd_detail
5. Howdy (2005) (in Japanese), <http://www.himanainu.jp/himag/?p=172>
6. Patel, S., Abowd, G.: The contextcam: Automated point of capture video annotation. In: Davies, N., Mynatt, E.D., Siio, I. (eds.) UbiComp 2004. LNCS, vol. 3205, pp. 301–318. Springer, Heidelberg (2004)
7. Watanabe, K., Tsukada, K., Yasumrua, M.: Willicam: a digital camera visualizing users' interest. In: Extended Abstracts of CHI 2007, pp. 2747–2752. ACM Press, New York (2007)
8. Hakansson, M., Ljungblad, S., Holmquist, L.E.: Capturing the invisible: designing context-aware photography. In: Proceedings of DUX 2003, pp. 1–4. ACM Press, New York (2003)

TreasurePhone: Context-Sensitive User Data Protection on Mobile Phones

Julian Seifert¹, Alexander De Luca², Bettina Conradi², and Heinrich Hussmann²

¹ Bauhaus-University Weimar, Bauhausstr. 11, D-99423 Weimar, Germany

julian.seifert@uni-weimar.de

² University of Munich, Amalienstr. 19, D-80333 Munich, Germany

{firstname.lastname}@ifi.lmu.de

Abstract. Due to increased input and output capabilities, mobile phones hold many different kinds of (mostly private) data. The need for finer grained profiles and integrated data security on mobile phones has already been documented extensively (e.g. [1]). However, there are no appropriate concepts and implementations yet to handle and limit access to data on mobile phones. TreasurePhone has been designed to address this specific problem. It protects the users' mobile phone data based on their current context. Privacy protection is realized by *spheres*, which represent the users' context-specific need for privacy. That is, users can define which data and services are accessible in which sphere. TreasurePhone exploits context information to support authentication and automatic activation of spheres by *locations* and *actions*. We conducted a user study with 20 participants to gain insights on how well users accept such a concept. One of the main goals was to find out whether such privacy features are appreciated by the users even though they make interaction slower and might hinder fast access to specific data. Additionally, we showed that integration of context information significantly increases ease-of-use of the system.

1 Introduction

Modern mobile phones support the creation and storage of many kinds of data ranging from contacts and e-mail to photos and text documents. At the same time, the amount of stored data is growing enormously which increases the need for securing the privacy of this data [2]. For instance, the integration of mobile phones into enterprise environments for mobile handling of e-mail, contacts and other data is enjoying increasing popularity. However, mobile phones still use a simple privacy/security model that only distinguishes between *locked* and *unlocked* [1].

Users have different contexts in their life such as family and work each with a corresponding need for privacy [3]. This makes privacy management of the data stored on their mobile phones practically impossible. That is, a user who has a single mobile phone for her working context as well as for private use cannot hide data belonging to one context while being in the other one. When working for companies that have high security standards, a user might face additional usage restrictions to avoid exposing business data to third parties by using the business mobile phone for private use as well.

One solution for this challenge would be to use more than one mobile phone. Users might have a mobile phone for their work as well as a personal one. From a usability

perspective this solution is not satisfying as there are usually more contexts than only *work* and *personal*. Therefore, users would need to use one mobile phone for each context they have.

We argue that privacy protection should be an essential part of the mobile device's operating system and should be addressed during the design of mobile systems. In this paper, we present TreasurePhone which supports context-sensitive protection of the user's data by allowing the user to define so called *spheres*. TreasurePhone uses *locations* for automatic activation of spheres and supports interaction with the user's environment to activate appropriate spheres on the go. TreasurePhone enables users to secure their data in each context in a sophisticated way using one mobile phone. Hence, TreasurePhone reduces the risk of unwillingly disclosing sensitive and private data.

2 Related Work

Work related to TreasurePhone can be generally classified into three categories: conceptual work about data privacy for mobile devices, authentication mechanisms for cell phones, and context-dependent adaptive mobile devices.

Stajano addresses privacy issues that arise from sharing (willingly or unintended) a personal digital assistant (PDA) with others [4]. He describes a system for PDAs which is based on the observation that some data and applications could be used by anybody who gets access to the PDA. However, other applications and data should be accessible only by the legitimate owner of the device. Accessing these *private areas* or "hats" would require authentication and thus secures the privacy of the user. In their work, Karlson et al. conducted interviews to find out basic requirements of data privacy on mobile phones. Their results suggest to use *usage profiles* that correspond to different contexts of the user [1]. These would allow sharing the mobile phone to others without risking disclosure of private data. They showed that users would appreciate a security model for mobile phones that is based on usage profiles enabling privacy management. However, the concept of usage profiles was not implemented. Nevertheless, this work, suggesting a role based access model, strongly influenced the design of TreasurePhone.

With *SenSay* Siewiorek et al. present a mobile phone that adapts its behavior in a context-based way [5]. This system processes data captured by several sensors and determines the user's current context based on the results. SenSay adapts the ringer volume, vibration and alerts to the current context. It can further provide remote callers with the ability to communicate the importance of their call which optimizes the availability of the user. Another contribution with its focus on context-based adaptation is presented by Krishnamurthy et al. [6]. Instead of using various sensors to determine the current context of a user, this system makes use of near field communication (NFC). With NFC, the context can be determined on a fine grained base. This system as well as SenSay manage to determine the context of the user, but use a different approach. Both systems do not focus on privacy issues or data security.

TreasurePhone provides a first implementation of a usage profile based system for mobile devices as suggested by Stajano and Karlson et al. The prototype applies findings presented by Krishnamurthy and Siewiorek and combines them to provide an advanced security model.

3 TreasurePhone

Threat model. In this work, we model two main threats against which the described system is resistant:

The first threat consists in unwillingly disclosing private or unappropriate data to the “wrong” people. Mobile phones are often borrowed to friends and other people, mostly to help them by providing a possibility to make phone calls, browse the Internet, etc. While interacting with the phone, the borrower might accidentally gain access to data that the owner of the mobile phone might want to keep private (e.g. when browsing the photos on the mobile device). Using TreasurePhone, a special sphere could be used that grants access to the call application only to avoid such problems.

The second threat are attackers that willingly try to steal information (e.g. important business data) from a user. By disabling (and encrypting¹) data of other contexts, TreasurePhone limits those kind of attacks. For instance, business data can only be stolen while the device is set to the business sphere.

Concept. Privacy cannot be seen as a fixed state. It rather means dynamically controlling the disclosure and use of personal information [7]. The dynamic character of privacy is stressed by its context-depended nature [3]. Furthermore, the user’s grasp of what kind of personal data is considered as private is highly individual [8]. In the field of sociology and psychology, the concept of *faces* exists that was proposed by Goffman [9]. According to Goffman, people use different faces depending on their current context; a face defines what information a person reveals to a specific audience.

The concept of TreasurePhone is based on the hypothesis that users are willing to protect and manage the privacy of their private data stored on their mobile phones. Based on Goffman’s faces we propose the concept of *spheres* that allow users to protect their data privacy. A sphere represents the user’s privacy requirements for data on her mobile phone in a specific context. That is, the user can define which applications such as e-mail clients, address books, photo viewers etc. are available in a specific sphere and furthermore, what exact data is accessible and which is not. One can imagine a sphere as a filter that lets pass only data that are not private in this sphere. This way, users could create spheres for their home, family and friends as well as work context – each providing only as much access to data as desired. The spheres concept includes one special sphere that allows exclusive administrative actions such as creating, editing or deleting spheres as well as deleting or changing access rights of data. This sphere is called *Admin Sphere* (AS) and requires the user to authenticate before accessing it. Usually this sphere will only be active when the user wants to perform administrative work. All other spheres do not allow deleting data or editing access rights of data. Besides the AS, TreasurePhone contains three spheres by default: *Home*, *Work* and *Closed*, which serve as examples of typical configurations that are not bound to certain contexts but can be applied in various matching situations. While *Home* provides access to all services, *Closed* denies access to all of them. This set of default spheres was compiled based on the results of a small study with five participants who used diaries to collect the contexts for which they would use spheres.

¹ Please note that this feature has not been implemented in the prototype.

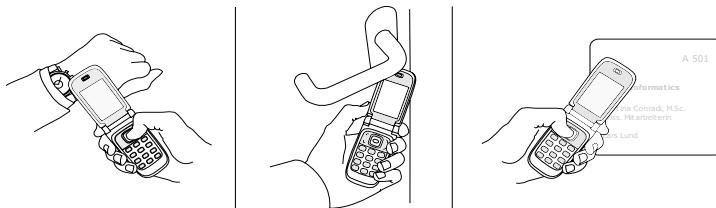


Fig. 1. a) Authentication using a personal token that is integrated into a wristband. b) Controlling a lock using actions. c) Reading a location that is based on an NFC tag integrated in a nameplate.

In order to protect the data, the user chooses the appropriate sphere depending on the current context. However, to prevent any person other than the legitimate owner from accessing private data, the activation of other spheres requires the user to authenticate to the system if the current sphere is not the *AS*. Fast and secure methods for authentication that do not require manual entry of a PIN minimize the effort for the user [4]. The TreasurePhone prototype supports authentication using a personal token that contains an NFC tag (see Figure 1a). It has to be noted here, that the benefit of the personal token comes with a security flaw. If an attacker can steal both, the token and the mobile device, full access to the device will be granted. To minimize the effort of spheres even further, context-dependent activation of spheres by *location* is supported by the system. A location in TreasurePhone is a configuration that is associated with a sensor value such as GPS coordinates, a Wi-Fi network identifier, a Bluetooth identifier or an RFID tag (see Figure 1c). Whenever a location is recognized, the corresponding sphere is activated. Besides locations, TreasurePhone supports interaction with the user's environment by *actions*. An example could be a Metro Network (like the Tokyo Metro system) that supports the use of NFC-enabled mobile phones to handle payment. When a user leaves the metro network at his work place, touching the gate mechanism with the phone would activate the *Work* sphere. Entering the metro network at his work location on the other hand could switch back to the *Closed* sphere.

Example Scenario. Using TreasurePhone implies initial effort for configuring the system. However, this is not mandatory because of the set of default spheres that are available. The configuration effort consists of creating individual spheres according to the user's needs and contexts in addition to the default spheres. For example, Bob could create a new sphere named *Friends*, which he intends to use while he is with friends, for instance at home or in a pub. He configures this sphere to allow access to messages, the address book and the photo service. Now Bob can start to create and manage data. After a while the configuration of Bob's TreasurePhone looks like the illustration in Figure 2. In the spheres *Home*, *Friends* and *Work* some contacts and other documents are visible. The spheres *Friends* and *Home* overlap and both allow access to the data in the intersection. The *Admin Sphere* encloses all data and Bob can access all data while this sphere is active.

When Bob turns on his mobile phone the *AS* is initially activated. After checking if there are new messages and having a look at today's appointments at work, Bob activates the *Home* sphere. Thereby personal data like photos, messages and contacts

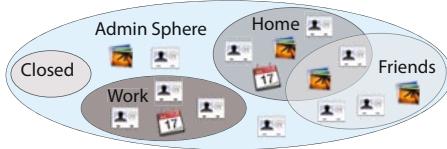


Fig. 2. The sphere model: The *Admin Sphere* allows access to all data; other spheres limit access and might overlap

are accessible, however, all business related data are hidden now. When Bob leaves his apartment he locks the RFID based lock of the door using his TreasurePhone, which is also usable as a key (See Figure 1b). This requires the configuration of corresponding *actions* for the lock. Bob configured the action *Locking Door* to activate the *Closed* sphere when finished. By using this action Bob does not have to think of changing the sphere. As Bob arrives at his office, his mobile phone detects the Bluetooth identifier of his desktop computer, which is associated with the location *My Office*. The sphere *Work* gets activated automatically. Now Bob has access to his calendars, documents, messages and all other data that is work-related. However, photos of his family and friends are now hidden.

Prototype Implementation. The TreasurePhone prototype is written in Java ME and implements the fundamental concepts: spheres, locations, actions and services as well as an abstraction for data. A sphere management subsystem controls which sphere is activated and what data and services are accessible. Activation is based on context information such as sensor data that correspond to locations and actions. The implementation also contains interfaces for applications which allows access management of applications that are registered as services.

The TreasurePhone prototype provides basic functionalities of standard mobile phones such as call, SMS, address book, camera, and a photo viewer. The user interface changes or grants access depending on whether the AS or another sphere is activated (see Figure 3). Editing access rights for data is only available while the AS is activated.

The default assignment of data access rights follows the basic rule: data is accessible in the sphere in which it was created. For instance, if the sphere *Home* is activated while



Fig. 3. Screens of TreasurePhone (AS activated): *a*) Editing access rights for a photo. *b*) Creating a new sphere named “Friends”. *c*) Editing contact details.

the user makes a photo, this picture is accessible by default in this sphere. In case of the *AS* being activated, the image would not be accessible in any of the normal spheres.

We chose the Nokia 6131 NFC mobile phone as platform for the first prototype, which comes with a built-in NFC reader. The prototype allows the user to authenticate via a personal token, which contains an NFC tag or by entering a PIN. NFC is also used for locations. The physical correspondence of a location in TreasurePhone is an NFC tag attached to an object (see Figure 1c).

4 User Study

We conducted a preliminary evaluation of TreasurePhone to study two basic questions. First, will users accept the increased complexity of handling the mobile device required by the privacy features? Second, will the use of automatic sphere switching by context (locations and actions) have a positive effect on the usability of the system? We randomly recruited 20 volunteers; 8 female and 12 male. Participants were undergraduate and PhD students with a technical background and aged between 23 and 32 years. They indicated they had all used mobile phones for at least six years. Half of the subjects use profiles (like silent, vibrate etc.) of their mobile phone on a daily basis; the others only occasionally or not at all. 19 of the subjects use PIN authentication when they turn on their mobile phone while only 3 use PIN authentication after each period of inactivity. During the study we first explained the system and then a training phase with the prototype was conducted by the participants. For training, each feature of the system was explained to them and tested with a small task. Next, practical tasks were carried out. Finally the users filled out a questionnaire regarding the system. Answers were given on a five point Likert scale (1=worst, 5=best). Overall the procedure took around 40 minutes, up to one hour.

The practical tasks started with a system configuration, in which users had to create and configure a sphere. This was followed by a series of five tasks in randomized order, which covered all actions that are specific for the concepts of TreasurePhone (see figure 4). For instance, participants created a contact in the address book and set the access rights for this contact to 'visible in sphere x'. Other tasks required the participant to activate different spheres in order to hide or get access to data. These five tasks were repeated two times in randomized order. One time participants used a prototype that did not integrate context information and a second time they used a system that supported context information integration. That is, one time the participants could make use of token based authentication (a wristband with an integrated NFC transponder), locations, and actions and the other time they could not. The context free prototype used an assigned PIN to activate the *Admin Sphere* and to switch between spheres.

Results of the study show that on average, users consider the system easy to understand ($\text{Avg}=4.4$, $\text{Mdn}=4$, $\text{SD}=.5$). They appreciate the support given by integrated context and 19 out of 20 participants stated that they would prefer using a system that implements locations, actions, and token based authentication. Users rated the general system's capabilities to secure privacy as 4.2 ($\text{Mdn}=4$, $\text{SD}=.8$) and the usefulness of spheres for privacy protection as 4.6 ($\text{Mdn}=5$, $\text{SD}=.5$). However, users estimated their willingness to store more sensitive data on their mobile phone, if this was running TreasurePhone, with 3.2 ($\text{Mdn}=3$, $\text{SD}=1.1$). Nevertheless, users stated that on average (4.1)

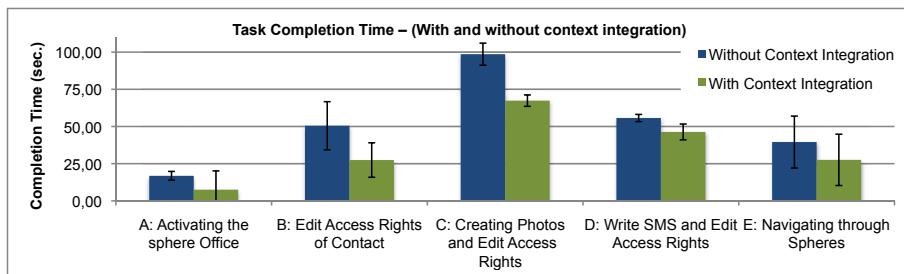


Fig. 4. Task completion times of the practical tasks with and without context information integration (error bars display the standard deviation)

they would feel more secure when sharing their TreasurePhone secured mobile phone with others ($Mdn=5, SD=1$).

Because this is a laboratory experiment, our results should be handled with care. However, they suggest user acceptance of the security features, and a preference for the context integration. Users did not mind increased complexity (and even did not consider it that complex). Also they agreed that their data would be more secure on such a phone. One user confirmed this by stating “I wouldn’t need to be concerned about my data so much when I want to share my mobile phone with a friend or when I just leave it at some place”. One user was especially happy that this system would provide her the possibility to limit the access to specific applications as well: “I like that I can even define access policies for facilities such as camera and address book”. The results are already quite encouraging, even more since none of the participants was in a business that requires carrying around sensitive data on a mobile device. We expect business users to be even more concerned about their data privacy.

A detailed analysis of task completion times shows that, not surprisingly, tasks were completed significantly faster with the prototype that uses context information for task switching (see Figure 4). The data was analyzed using paired-samples t-tests. For each task the prototype using NFC was faster than the PIN version. The results for task A ($t(18)=7.26, p<.001$), B ($t(16)=4.15, p<.003$), C ($t(15)=5.91, p<.001$) and D ($t(18)=3.85, p<.003$) were highly significant while the difference in task E was significant ($t(17)=2.89, p<.05$). The positive results for the context version are supported by the users’ opinion. One user explicitly stated “it makes changing the profiles fast and easy”.

5 Conclusions and Future Work

In this work, we presented TreasurePhone, an approach toward a mobile phone operating system which supports context dependent data privacy for users based on spheres. Supporting locations and actions for changing spheres makes adapting to the users’ current context easier. The results of the user study show that integrating context and fast authentication makes the system significantly faster in use and is favored by the users over a system that requires manual authentication and manual sphere switching.

While our study suggests that users are interested in the security and privacy provided by TreasurePhone, future studies of long term use would be valuable to determine whether users prefer using spheres to existing “binary” security models in day to day use of their phones. Steps toward answering this question include implementing an advanced prototype, whereas spheres are integrated at the operating system level in order to meet the requirements for a longterm study. Additionally, we would like to implement support of further sensors for interaction with locations such as GPS or Bluetooth identifiers and thus extend TreasurePhone’s context sensitivity. A very interesting aspect with respect to the sensors is which of them are actually suitable for context switching from a usability’s point of view. That is, which of them can be used and understood by the users.

Acknowledgments

We would like to thank the reviewers and especially A.J. Brush for their fruitful comments that greatly helped to improve this paper.

References

1. Karlson, A.K., Brush, A.J.B., Schechter, S.: Can I Borrow Your Phone?: Understanding Concerns when Sharing Mobile Phones. In: CHI 2009: Proceedings of the 27th international conference on Human factors in computing systems (2009)
2. Stajano, F.: Will Your Digital Butlers Betray You? In: WPES 2004: Proceedings of the 2004 ACM workshop on Privacy in the electronic society. ACM, New York (2004)
3. Lehikoinen, J.T., Lehikoinen, J., Huuskonen, P.: Understanding privacy regulation in ubicomp interactions. Personal Ubiquitous Comput. 12(8), 543–553 (2008)
4. Stajano, F.: One user, many hats; and, sometimes, no hat - towards a secure yet usable pda. In: 12th Int. Security Protocols Workshop. Springer, Heidelberg (2004)
5. Siewiorek, D., Smailagic, A., Furukawa, J., Krause, A., Moraveji, N., Reiger, K., Shaffer, J., Wong, F.L.: SenSay: A Context-Aware Mobile Phone. In: ISWC 2003: Proceedings of the 7th IEEE International Symposium on Wearable Computers, Washington, DC, USA. IEEE Computer Society, Los Alamitos (2003)
6. Krishnamurthy, S., Chakraborty, D., Jindal, S., Mittal, S.: Context-Based Adaptation of Mobile Phones Using Near-Field Communication. In: Annual International Conference on Mobile and Ubiquitous Systems, pp. 1–10 (2006)
7. Jiang, X., Hong, J.I., Landay, J.A.: Approximate Information Flows: Socially-Based Modeling of Privacy in Ubiquitous Computing. In: Borriello, G., Holmquist, L.E. (eds.) UbiComp 2002. LNCS, vol. 2498, p. 176. Springer, Heidelberg (2002)
8. De Luca, A., Hußmann, H.: Threat Awareness - Social Impacts of Privacy Aware Ubiquitous Computing. In: INTER: A European Cultural Studies Conference in Sweden (INTER 2007), Norrköping, Sweden, June 2007, pp. 1650–3686 (2007)
9. Goffman, E.: The Presentation of Self in Everyday Life. Doubleday Anchor Books, New York (1959)

Recruitment Framework for Participatory Sensing Data Collections

Sasank Reddy, Deborah Estrin, and Mani Srivastava

Center for Embedded Networked Sensing
University of California at Los Angeles, USA
sasank@ee.ucla.edu, destrin@cs.ucla.edu, mbs@ee.ucla.edu

Abstract. Mobile phones have evolved from devices that are just used for voice and text communication to platforms that are able to capture and transmit a range of data types (image, audio, and location). The adoption of these increasingly capable devices by society has enabled a potentially pervasive sensing paradigm - participatory sensing. A coordinated participatory sensing system engages individuals carrying mobile phones to explore phenomena of interest using *in situ* data collection. For participatory sensing to succeed, several technical challenges need to be solved. In this paper, we discuss one particular issue: developing a recruitment framework to enable organizers to identify well-suited participants for data collections based on geographic and temporal availability as well as participation habits. This recruitment system is evaluated through a series of pilot data collections where volunteers explored sustainable processes on a university campus.

Keywords: Mobile Computing, Participatory Sensing, Urban Sensing.

1 Introduction

The recent proliferation of mobile smart phones combined with the ease of deployment of web services for storage, processing and visualization, has ushered in a new pervasive data collection model - participatory sensing [123]. By enabling people to investigate previously difficult to observe processes with devices they use everyday, participatory sensing brings the ideals of traditional community based data collection and citizen science to an online and mobile environment; offering automation, scalability, and real-time processing and feedback [45]. In participatory sensing, individuals explicitly select the sensing modalities to use and what data to contribute to larger data collection efforts. Example initiatives that are enabled by participatory sensing include our pilot data collections, where individuals collected photos of assets that documented recycling behavior, flora variety, and green resources to learn more about sustainability at a university.

However, advancing participatory sensing from a potential to a coordinated reality remains a major challenge. Finding a fit between diverse users and participatory sensing projects mirrors traditional selection for volunteer work based on interest and skill. But because participatory sensing is organized virtually,

identifying particular participants (individuals who collect, analyze, and share their data) for campaigns (targeted data collection efforts) can be partially automated. Identification can rely not only on participants' reputations as data collectors based contribution habits, but can also be enhanced by incorporating participants' availability in the area of interest [6][7][8]. Specific attention is payed to the fact that humans have self-will, exhibit varied data collection performance, and have mobility traits that are opportunistic in nature [9].

This paper proposes a recruitment framework for participatory sensing data collections. Our work makes the following contributions: (a.) identifies availability and data collection performance as core attributes needed to match participants to campaigns, (b.) details models and algorithms that can be used to represent the recruitment factors, and (c.) evaluates the usefulness of the proposed recruitment mechanisms through pilot data collections. The rest of the paper is organized as follows: Section 2 illustrates example campaigns and motivates the recruitment problem. Section 3 provides an overview of the approach taken to address the recruitment challenge. Section 4 describes related work, and system details are given in Section 5. The paper ends with an evaluation of the recruitment framework and a discussion passage in Section 6 and 7 respectively.

2 Motivation and Application Examples

The application area for our data collections was an effort to learn more about sustainability practices at a university. A series of campaigns that documented various resource use issues were initiated. The data collections were enabled by a system consisting of a mobile phone client (Android G1 and Nokia n95) along with web services for data storage (Flickr and sensor database), analysis (Python application server), and visualization (Google Maps and Charts). Figure II a.) contains the mobile phone and web feedback page user interfaces. The campaigns involved taking geo-tagged photos, Figure II b.), and are described below:

- **GarbageWatch:** The campus needs to divert 75% of its waste stream from landfills, and effective recycling can help reach this goal. Participants documented the contents of outdoor waste bins through photo documentation. By analyzing the images, one can determine if recyclables (paper, plastic, glass, or aluminum) are being disposed of in waste bins, and then identify regions and time periods with low recycling rates.
- **What's Bloomin:** Water conservation is a high priority issue for the campus and efficient landscaping can help. This campaign involved taking geo-tagged photos of “blooming” flora. Having this inventory enables facilities to replace high water usage plants with ones that are drought tolerant. This flora catalog does not exist since the landscape is managed by many groups.
- **AssetLog:** For sustainable practices to thrive on a campus, the existence and locations of “green” resources needs to be documented. These resources include bicycle racks, recycle bins, and charge stations. But with expansion and re-construction activities, an up to date list is not available. Thus, this campaign tasked individuals to capture photos of these sustainability assets.



Fig. 1. System User Interface Design and Campaign Image Examples

Participatory sensing campaigns seek individuals willing to collect data about a particular phenomenon. A recruitment service takes campaign specifications as input and recommends participants for involvement in data collections. Campaign specifications may involve a number of factors including participants' device capabilities, demographic diversity, and social network affiliation. However, this work concentrates on a specific set of requirements for recruitment: participants' reputations as data collectors and availability in terms of geographic and temporal coverage. Also, our campaigns have an overall budget associated with them which may include resources needed to run the data collections along with compensation when incentives are provided for participant involvement. In our system, reputation is limited to considering participants' willingness (given the opportunity, is data collected) and diligence in collecting samples (timeliness, relevance and quality of data). Availability is learned from previously collected context-annotated mobility traces (i.e. streams of location, time, and transportation mode) in the campaign coverage area. Thus, the recruitment step would be used by campaign organizers to select participants who achieve the highest data collection utility while adhering to the set campaign budget. Overall, our recruitment framework is best suited for campaigns that have systematically defined data collection guidelines and are constrained in terms of coverage.

The sustainability campaigns are used to illustrate the features of the recruitment system. For these campaigns, well-suited participants are ones that regularly walk on campus during daytime hours and cover as much of the campus area as possible. Individuals that run, bike, or drive may be less likely to notice the resources of interest, and collecting clear photos is difficult at night. Furthermore, it is important that participants are willing to make observations when given the opportunity and that these samples are relevant and high quality.

3 System Overview

The process of recruiting volunteers for participatory sensing campaigns is analogous to recruiting volunteers or employees in non-virtual environments. Drawing on this similarity, we have created a recruitment framework, illustrated by Figure 2, that consists of three stages: the qualifier, assessment, and progress review.

- **The Qualifier:** Participants for campaigns must meet minimum requirements. For availability, prerequisites are based on destinations and routes within time, space, and transportation mode constraints. For participation reputation, requirements are measures of sampling likelihood, quality, and validity over several campaigns or by campaign-specific calibration exercises.
- **The Assessment:** Once participants that meet minimum requirements are found, the recruitment system then identifies which subset of individuals maximize coverage over a specific area and time period while adhering to the required transportation modes. Participants have costs and there exists a campaign budget which are both considered when selecting participants.
- **The Progress Review:** As a campaign runs, the recruitment system must check participants' coverage and data collection reputation to determine if they are consistent with their base profile. This check can occur periodically, and if the similarity of profiles is below a threshold, organizers should be alerted so that they can provide feedback or recruit additional participants.

The design of the recruitment system takes into account the private nature of availability and participation data. Thus, the three-stage framework works to be parsimonious by limiting both the amount and granularity of information that is shared. Also, our system is designed to be run in coordination with a personal data vault where all participant information is stored and external queries on this data are strictly opt-in [10][11]. For the qualifier and progress review, the query results sent to the data vault will simply be aggregate results of whether conditions or thresholds are met. In the case of the assessment, more detailed data in regards to mobility profiles needs to be shared with the recruitment system since coverage is based on collective participant mobility, but the data is limited to a particular spatial region, time span, and transportation mode.

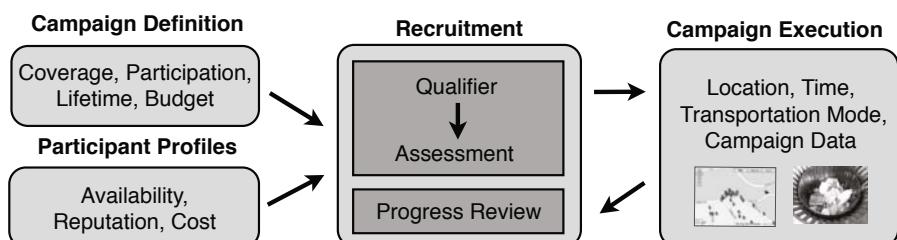


Fig. 2. Recruitment Framework Inputs, Outputs, and Steps

4 Related Work

An overview of related work in terms of models, algorithms, and systems that share properties similar to the participatory sensing recruitment system is provided. First, models used to represent mobility and reputation that exist are reviewed. Then, details about systems that share a similar purpose of selecting resources for a task based on set conditions are documented.

4.1 Mobility Models

Location Summarization for Personal Analytics. There has been a significant amount of work in regards to coming up with clustering algorithms to summarize the most significant destinations of a user based on location traces [12][13][14]. The location traces can come from a GPS receiver, access point (GSM or WiFi) mappings, or hybrid setups that combine GSM, WiFi, and GPS. To derive the significant destinations, consecutive location points within a certain time period are aggregated into clusters. Also, certain systems use map matching and reverse geo-coding to add additional contextual information (semantic meaning) to the clusters [15]. This information has been used to create “gazetteers” (geographical dictionaries) for individuals. In terms of the recruitment framework, the qualifier step will have to use a similar clustering scheme as these systems since the granularity of summarization is on the destination and route level.

Location Prediction to Adapt Applications. Mobile quality of service (QoS) and location based services (LBS) have used location prediction in order to improve and enable applications. The mobile QoS work mainly concentrates on creating systems that provide predictive and adaptive bandwidth reservation for mobile phone users based on their short term mobility. These models take a very microscopic view on mobility, concentrating on determining which “cells” a user might travel based on transition patterns from previous cells, time spent in the current cell, and speed/trajecotry information [16][17]. Most LBS use the current location of a user for application adaptation, for instance in traffic, entertainment, and shopping settings. But researchers have proposed to make LBS more relevant to the user’s next destination. For instance, [18] models transitions of individuals using Markov processes, and [19] incorporates factors such as land-use information to help with destination prediction. Although the recruitment service does not require this type of short term prediction, the underlying algorithms to model historical location data is relevant.

Mobility Based Networking. The Mobile AdHoc Networking (MANET) community has used mobility models that simulate the movements of individuals to test out performance of networking protocols [20][21]. Early work concentrated on using a random waypoint model where a node is specified with certain speed, direction, and duration of travel and then simulated to generate mobility patterns by randomly changing these factors after a period of time. Recently, these

models have gotten more sophisticated with the inclusion of geographical constraints and historical information, but they are still mainly useful for generating statistically equivalent traces and not for modeling existing real world traces.

Delay tolerant networking (DTN) has also used mobility models in order to manage routing of messages so that systems would work in situations that do not have continuous network connectivity. These systems rely on creating location matrices that model the presence of an individual at different locations and then compare the profiles of users to figure out where to disseminate a message so that it will eventually end up in the target location [22][23]. Similar type of modeling has been used by the Reality Mining project to learn about location habits of groups [24]. Overall, this work is relevant to the project review step in the recruitment process since mobility profile similarity checks are needed.

4.2 Reputation Models

Summation and Average. The simplest reputation models are ones that are summation and average based. In this setup, ratings are aggregated, by summing or averaging, to create an overall single reputation score [25]. An example of a summation system is eBay where ratings, which can be either -1 (negative), 0 (neutral), and 1 (positive), are added together [26]. Amazon instead uses averaging and relies on a “star” rating system that ranges from 1-5 where 1 is poor and 5 is excellent [27]. The advantage of these models is that they are easy to understand since a single number represents reputation, but the disadvantage is that they provide a primitive view on an individual’s actions and can cover up negative ratings if many positive ratings exist in proportion [28].

Discrete Trust Models. An alternative scheme to having reputations being a numerical value is to use discrete labels. For example, the Slashdot web site aggregates ratings on actions, such as story submissions, postings, moderation activities, into “karma” tiers for participants that include terrible, bad, neutral, positive, good, and excellent [29]. Although this model is helpful for individuals to quickly determine a meaning for a reputation measure, it is not mathematically tractable and has no method to determine reputation confidence [28].

Bayesian Systems. Reputation models based on Bayesian frameworks have been popular for peer-to-peer networks and sensor systems [25][30]. Particularly, these systems rely on ratings, either positive or negative, and use probability distributions, such as the Beta distribution, to come up with reputation scores [31]. By taking the expectation of the distribution, reputation can be determined. The confidence in this reputation score is captured by analyzing the probability that the expectation lies within an acceptable level of error. Additional features are easily enabled, such as aging out old ratings by using a weight factor when updating reputation and dealing with continuous ratings by employing an extension involving the Dirichlet process [30][31]. Overall, this Bayesian framework, specifically with the Beta distribution, seems to be appropriate to model participant data collection habits.

4.3 Selection Services

Crowd-Sourcing Sites. Many crowdsourcing services on the web have requirements that need to be met before individuals can take part in a task [32]. Sites like Amazon Mechanical Turk and GURU.com, which are systems that provide a marketplace to get commissioned work done, keep detailed statistics tracking the performance of workers. In Amazon Mechanical Turk, work done by a participant is evaluated in terms of whether it was accepted or rejected by requesters [33]. In GURU.com, the technical skill, creativity, timeliness, and communication capabilities of a worker are kept through a star-based rating system based on feedback from work requesters [34]. Our work builds on this idea of monitoring user behavior and provides metrics to evaluate participation and performance of individuals involved in data collection.

Sensing Systems. Sensor network research has taken place in regards to selecting and placing static devices to maximize coverage [35]. Similarly, work exists to coordinate robotic motion for sensing purposes [36]. Unfortunately, the algorithms for these systems do not apply directly to the recruitment problem since mobility of individuals is not always controllable and there exists variability in when and how sampling occurs. Previous work related to mobile phone opportunistic sensing either concentrate on creating protocols to recognize when sensing should be activated based on pre-defined zones [6][7][37] or choosing how much sensing should occur depending on privacy restrictions [38]. Our work differs in that the data collection recruitment problem is directly addressed with participant availability, reputation, and coverage/participation inconsistency considered. Also, our system does not rely on knowing prior distribution information or having detailed statistical models of the phenomenon of interest.

5 System Details

The steps involved in both availability as well as participation and performance based recruitment are detailed below. Specifically, we focus on the inputs and outputs of each of the different steps in the recruitment framework. Also, we detail the models and algorithms involved in the framework.

5.1 Coverage Based Recruitment

Mobility Information. Coverage based recruitment relies on transforming raw participant mobility data into building blocks that can be used for processing. The system assumes that participants have previously collected location traces in the form of latitude, longitude, and time points for a period of time that represent their “typical” behavior (e.g. for a profile week). The location traces could be augmented with sensor-based information which can help in adding context such as transportation mode (still, walking, running, biking, or driving) [39][40][41]. Having this type of data collected by participants is not far fetched; services already exist that rely on location check-ins and traces [42][43].

Qualifier. The transportation mode annotated location traces are transformed into significant destinations and routes for the qualifier. The system pre-processes the data by normalizing it to a set sample rate (for instance, every 30 seconds) and fills in missing values when the GPS signal is lost. For large spatial gaps, the points are filled by generating likely traces using the Google Maps API. Then, location points within a certain time period (at least 15 minutes) and distance bound (50 meters) are grouped into “stays” [44]. Density based clustering is used to group stays within a certain distance (250 meters) into “destinations” [44]. Routes are points between destinations and are aggregated using hierarchical clustering where the average minimum point segment distance is the comparison metric [45]. Qualifier queries use these building blocks to create filters, such as participants that have at least 5 destinations in a certain area in a week or individuals that have 7 unique walking routes during day time weekday hours.

Assessment. Next, in the assessment step a subset of qualified individuals that maximize coverage are identified. Formally, the assessment is an instance of the budgeted maximum coverage problem [46]. A participant pool, $P = \{p_1, p_2, \dots, p_n\}$, exists with non-negative costs, $\{c_i\}_{i=1}^n$. Spatial and temporal blocks with an associated transportation mode, $E = \{e_1, e_2, \dots, e_n\}$, are present. The blocks have utilities, $\{u_i\}_{i=1}^n$, defined for the campaign as well. The goal is to find a subset of participants $P^* \subseteq P$, such that the utility of elements covered, $U(P^*)$, is maximized while the cost of the subset, $C(P^*)$, is under a set campaign budget, B [38][46]. Hence, the optimization can be stated as:

$$\text{argmax } U(P^*) \text{ subject to } C(P^*) \leq B \quad (1)$$

This optimization is NP-hard since selecting a participant for the subset changes the utility for the rest not included. Thus, to find the best solution, all subset combinations must be searched. Since the utility function is sub-modular (adding a participant helps more if fewer are already selected) and non-decreasing (utility of subset is less than the set it is derived from), the greedy algorithm can find an adequate solution when costs are identical (at least 63% from the optimum) [47]. If costs are not identical, the benefit-cost greedy algorithm can be used where the ratio of utility to cost is used as the metric to pick participants [38]. Alternatively, this algorithm can help find the least costly subset to achieve a coverage goal.

Progress Review. While a campaign runs, check-ups are needed to ensure that participant mobility is consistent with the profile used for recruitment. Thus, in the progress review the similarity of mobility profiles is checked. To model mobility for the progress review, a time span (one week) is represented using an association matrix, A , consisting of $m \times n$ entries [22][48]. The m rows indicate spatial blocks (e.g. 10000 meter² grids) while the n columns model distinct time periods (days). An entry in the matrix is the proportion of time spent in a location performing set transportation modes within the time period selected. A day is chosen as the representative time period while a week is the time span based on previous work on human location patterns [24][49].

Since it is only necessary to compare the dominant mobility patterns, a summarization technique for the association matrix is needed. Thus, Singular Value Decomposition (SVD) is applied to the association matrix: $A = U \cdot \Sigma \cdot V^t$. In this decomposition, U , the left eigenvectors, are referred to as eigenbehaviors and represent patterns that are common across different time periods (days), and the singular values Σ represent the variance represented by each pattern. Consecutive time spans (weeks) are compared by taking the cosine similarity of the behavior vectors weighted by the singular value importance [22]. Hence, if there exists two eigenbehaviors, U_{t1} and U_{t2} , representing different time spans, $t1$ and $t2$, with singular value importance, W_{t1} and W_{t2} , the similarity metric is:

$$\text{Similarity}(U_{t1}, U_{t2}) = \sum_{i=1}^{\text{rank}(U_{t1})} \sum_{j=1}^{\text{rank}(U_{t2})} w_{t1_i} w_{t2_j} |U_{t1_i} \cdot U_{t2_j}| \quad (2)$$

Similarity is indexed from 0 (least similar) to 1 (most similar) by normalizing on the base eigenbehavior similarity.

5.2 Participation and Performance Based Recruitment

Inspired by reputation metrics in other domains (Section 4.2), we divide data collector reputation into two classes: cross-campaign and campaign-specific. Cross-campaign indicators, such as the number of campaigns volunteered, participated in, and abandoned, provide a granular view of a participant’s experience across many campaigns. Campaign-specific metrics measure the quality and quantity of samples that can be expected for a specific data collection. In our work, we concentrate on campaign-specific measures, specifically on participation likelihood. Other examples include timeliness, relevancy, and quality of samples.

Timeliness represents the latency between when a phenomenon is sampled (or occurs) and when it is available for analysis. It is influenced by user and upload delay. Relevancy indicates how well the sample describes the phenomenon of interest. It ranges from describing the item that is desired to not being related at all. Quality represents the ability of a processing module to determine a particular feature for further classification. Participation likelihood describes whether an individual took a sample when given the opportunity. These measures can be automatically quantified or might require human intervention. The campaign organizer defines a utility function that combines the importance of each metric to determine the overall reputation for a participant on a per campaign basis.

Modeling. The Beta distribution is adopted for campaign-specific reputation since it can be stored and updated efficiently, estimate stochastic (due to the randomness of the system) and epistemic uncertainty (due to lack of knowledge about the randomness of the system), and have features such as aging added on top easily. The distribution is indexed by alpha (α) and beta (β), which define the number of successful and unsuccessful events and is expressed as follows:

$$f(p|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1} \quad (3)$$

A participant’s reputation can be found by calculating the expectation of the Beta distribution (stochastic uncertainty), $E(\alpha, \beta) = \alpha / (\alpha + \beta)$. Confidence in this reputation score (epistemic uncertainty) is the posterior probability given the actual expectation value lies within an acceptable level of error found by calculating the area under the Beta curve [50]. Alpha and beta are set to 1 initially, which results in a uniform distribution where all values are considered equally likely. Distributions that represent more evidence for a hypothesis are peaked at the expectation compared to ones with less evidence. Also, if continuous ratings are needed an extension of Beta involving the Dirichlet process can be used [30].

Qualifier. Most campaigns will not have a prior participant reputation data as related to the specific data collection that is of concern. Thus, it is necessary to go through a “calibration” exercise so that evidence is gathered for the qualifier step. This exercise commonly involves having an expert gather ground truth on set paths and directing participants to traverse them as well. In cases where an expert cannot be involved, participants simply get compared against each other on these paths. The contributions are evaluated in the Beta framework and the qualifier step removes individuals that do not have a certain reputation level.

Progress Review. As a campaign runs, the participation and performance of individuals could change. For example, individuals might be initially very diligent about data collection but then change their behavior due to loss of interest or schedule tensions. Thus, it is important to be able to check reputation based on the most current information. The Beta distribution provides the ability to consider discounting old information by using an aging factor, w . This aging is done by discounting existing reputation values at set intervals when updates occur [31]. Essentially, alpha and beta are transformed as follows:

$$\alpha_{new} = w_{age} * \alpha_{old} + \alpha_{obtained}; \quad \beta_{new} = w_{age} * \beta_{old} + \beta_{obtained} \quad (4)$$

6 Evaluation

This section analyzes the models and algorithms involved in coverage and reputation based recruitment. The sustainability campaigns provide the data for the evaluation. Importance is placed on highlighting the features of the framework.

6.1 Campaign Deployment Information

The sustainability campaigns were initiated by engaging individuals from campus student groups. Individuals were given a phone, trained on what to identify, and how to use the data collection software. Participants ran a campaign for at least one week although many continued for additional days (results shown in Table 1). Before the campaigns started, all individuals performed calibration exercises where they would go on pre-defined routes to collect data. These routes were also traversed by “experts” who gathered ground truth. During the campaign, participants did not receive instructions on where and when to sample.

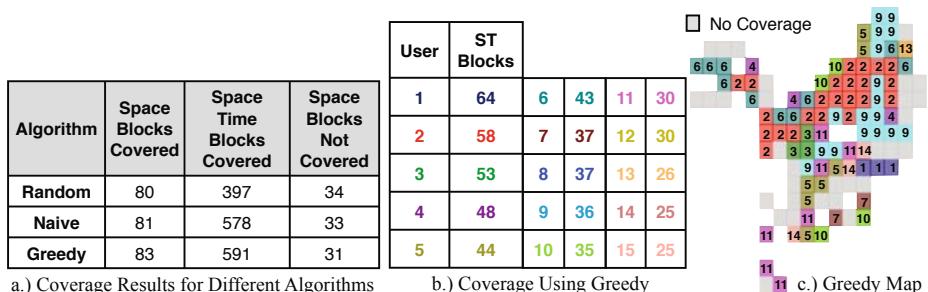
Table 1. Campaign Participation Data

Campaign Type	Total Images	Total Users	Average Per User	Maximum Per User	Minimum Per User
GarbageWatch	1752	31	56	231	7
What's Bloomin	4041	22	183	398	4
AssetLog	1488	16	93	266	11

6.2 Coverage Based Recruitment

The usefulness of coverage based recruitment is analyzed with the GarbageWatch campaign. Specifically, we focus on the assessment and progress review stages. Participants have already passed the minimum qualification of having routes and destinations on the campus since they all belonged to the university community.

Assessment: Evaluating the Best Coverage of the Campus. For GarbageWatch, the spatial zones of interest are campus waste bin locations, temporal span is daytime weekday hours, and the transportation mode is walking. The size of the spatial blocks was set to 10000 m^2 , which was empirically chosen based on GPS error and waste bin density, and the temporal block granularity was set to 1 hour so recycling behavior over time can be monitored. Three participant selection methods were compared: random, naive, and greedy. Random selects individuals for campaigns arbitrarily. Naive represents a heuristic where selecting participants is based on which individual covers the most blocks overall without considering what existing selected participants have covered. Greedy chooses participants that maximize utility while taking into consideration the coverage by existing selected participants. Thus, for greedy the participant utilities need to be re-calculated after an individual is selected. For evaluation purposes, the block utilities are all the same, the participant costs are set to 1, and the budget is limited to 15. In essence, 15 individuals are chosen from the pool of 31.

**Fig. 3.** Algorithm Comparison for GarbageWatch Campaign Coverage

The coverage results for the algorithms are shown in Figure 3 a.). The number of spatial temporal blocks is 6840 made up off 114 spatial blocks (based on 10000 m^2 granularity) and 60 time blocks (12 daytime hours per day for 5-weekday span). Furthermore, Figure 3 b.) shows specific coverage information for the greedy case, and Figure 3 c.) illustrates the greedy results on a map with the participant with the most coverage for a spatial block taking ownership. Random selection performs much worse then either the naive or greedy algorithms, specifically picking participants that have less spatial and temporal coverage and more spatial blocks not covered by anyone. The greedy algorithm performs better then just the naive heuristic. If more coverage overlap existed between participants, the performance of the greedy algorithm would be even higher. In general, considering availability when selecting participants is important. Otherwise, large coverage gaps could exist, and the opportunities available for sensing could be low. Also, the more complex instance of this problem, with variable costs for participants and different utilities for spatial and temporal blocks, can be handled by using a variant of the greedy algorithm where the benefit to cost ratio is used to evaluate participants during the selection process [38].

Progress Review: Comparing Coverage Profiles Over Time. As a campaign runs, participants availability might deviate from their established profiles. Thus, campaign organizers should be able to run checks on mobility profile consistency so that actions, such as recruiting additional individuals or providing feedback to existing participants, can take place if there is coverage loss. This progress review consistency check is especially important for long running campaigns since schedules might shift. The usefulness of the progress review is shown by analyzing two participants involved in the sustainability campaigns. One participant had a very stable schedule while the other had a significant shift occur. The mobility profile check is run by calculating similarity between eigenbehaviors of two weeks using SVD. Participants' mobility is modeled using an association matrix which is 114 (number of spatial blocks when considering a spatial granularity of 10000 m^2) by 5 (number of weekdays in a week) in size that takes into account daytime walking instances on campus during a week.

The mobility map and similarity score of Participant #9 is shown as Figure 4 a.), and based on interviewing the individual, we find that the participant mainly travels between two main hubs on campus and does not typically deviate. Thus, the similarity score of 0.85 based on comparing eigenbehaviors between the two weeks makes sense. In some cases, an individual might have a shift in schedule or a change in the way they travel. This was the case with Participant #2 who changed their transportation mode between their residence and campus from walking to driving between weeks. As shown in Figure 4 b.), this individual's similarity score is only 0.34. Overall, the SVD based similarity measure is effective to learn about major availability changes. Also this method has the advantage of summarizing mobility patterns in a compact manner - aggregating weeks of similar mobility data into a few dominant eigenbehaviors [8][22][48].

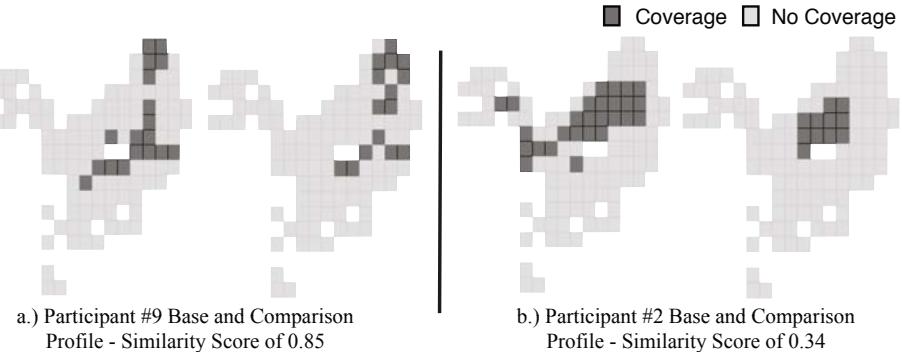


Fig. 4. Progress Review Consistency Check for Campaign Coverage

6.3 Participation and Performance Based Recruitment

Another factor to consider during recruitment is participants' reputations as data collectors. Although factors such as sample timeliness, relevancy, and quality can play a role in reputation, in our sustainability campaigns we found these elements to be less applicable since automatic image uploading was occurring, the items that needed to be sensed were distinctive, and participants took very few unusable images. Thus, we concentrate on whether a participant is likely to contribute a sample if they had an opportunity. This metric is used to exercise the features of the Beta distribution in the qualifier and progress review stages.

Qualifier: Running Calibration Exercise for Initial Reputation. Initially, for the three campaigns, no prior information existed in terms of sampling reputation. Thus, calibration exercises were implemented to get an initial sense of a participant's likelihood to capture a sample if they had a chance. This was done by having three specific routes that participants had to traverse for each campaign. Ground truth information was obtained along these paths by an "expert". For the case of GarbageWatch, opportunities to sample were places where waste bins existed. Similarly, for What's Bloomin the opportunities were related to places where flowers existed, and for AssetLog, each route was associated with a color and items of that color were samples of interest. The calibration routes were chosen to be paths that individuals on campus are familiar with. The calibration is run by participants once at the beginning of a campaign.

When designing a calibration exercise, an important factor to consider is whether there are enough sampling opportunities to be able to be confident of the reputation that is derived. For instance, in the case of the AssetLog campaign, if one route is only considered instead of all three to calculate initial reputation, then the campaign organizer might not have confidence in the reputation prediction provided. For example, Figure 5 shows Beta distributions for a participant where one route is compared to all three routes. As Figure 5 a.)

shows, even though the reputation of the participant is high with a score of 0.77 (likely to sample the phenomenon when given the chance), our confidence in his ability is low since the number of check points for sampling is small when considering only one route. When all three routes are used, Figure 5 b.), the confidence we have in the overall reputation of 0.81 is much higher. In fact, the confidence is at a level of 0.97 with all three routes considered as compared to just 0.61 when one is used. The confidence score was calculated by taking the area under the Beta curve with an acceptable error of 0.1 around the mean reputation.

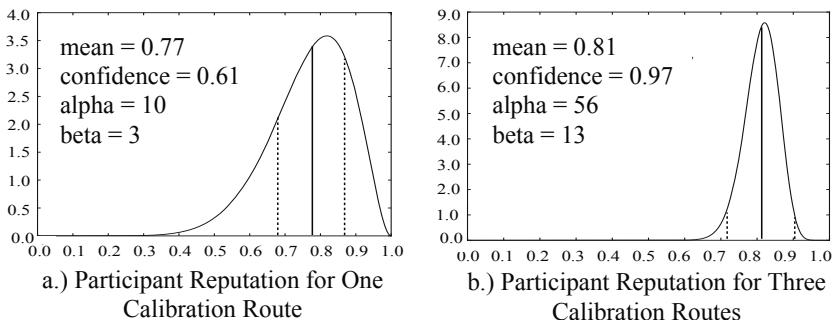


Fig. 5. Calibration Reputation for Participant in AssetLog Campaign

A question that comes up is whether these calibration exercises are useful as a predictor of sampling behavior during the actual campaign. To test this we compare the reputation gathered from the calibration exercises to the reputation derived when the participant ran the actual campaign. Since mobility traces were collected while the participants were performing the campaigns, we analyzed when they took images compared to when they had an opportunity. For GarbageWatch, prior information on all the waste bins locations existed and for the What's Bloomin and AssetLog campaigns, collective knowledge gathered from the participants submissions were used as ground truth. Table 2 shows the average of the percent difference of reputation for each participant in the three campaigns. The values are calculated by taking the difference between the calibration reputation and the reputation derived from the campaign and then averaging per campaign. The results, an average of 12.5% in reputation difference when considering all campaigns, indicate that the calibration exercises are reasonable approximations for participants real campaign reputations.

Table 2. Comparison of Calibration to Real Campaign Reputation

	GarbageWatch	What's Bloomin	AssetLog
Reputation Difference	10.3%	12.4%	14.8%

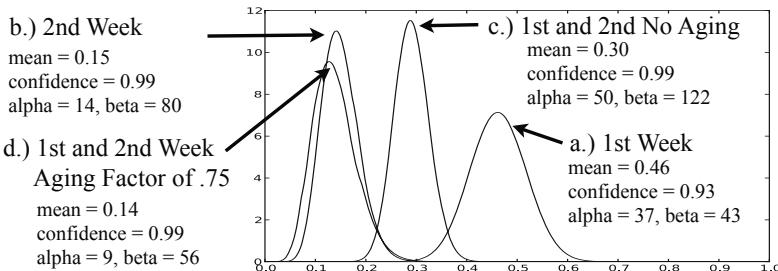


Fig. 6. Reputation for Participant Considering Aging Factor

Progress Review: Checking Reputation Over Time. Since there is a chance that sampling behavior could change as campaigns run, it is important to check participant reputations at set points as part of a progress review. Introducing aging on top of the Beta distribution can help with this checkup since it can be used to obtain a more current indication of an individual's reputation. We exercise this feature by analyzing the contributions of an individual that was involved in What's Bloomin for an additional week. Figure 6 shows the reputation, along with the Beta curves, of this participant based on their first week, second week, and then two methods to combine their weeks (with and without aging). During the first week, the participant's reputation to sample when given the opportunity was 0.46. But on the second week, their reputation is much lower at a level of 0.15. At the two week period, if all contributions were considered equally, the participant's reputation would be 0.30 but this is not indicative of the recent performance change. Instead if an aging factor of 0.75 (where 1.0 represents keeping all history and 0.0 is only considering the current information) is used to discount past reputation daily, then the end reputation is 0.14 which is a better indicator of the recent behavior shift.

7 Discussion

This section concludes the paper by summarizing lessons learned for campaign recruitment based on the evaluation results. Also, feedback provided by participants on their experience of performing campaigns is presented. Finally, future work that makes the recruitment system more flexible and adaptive is reviewed.

7.1 Recruitment Framework Analysis

The evaluation results reveal some important lessons learned for the recruitment framework. When analyzing the performance of the different algorithms during the assessment stage (Section 6.2), we find that selecting individuals based on using the greedy algorithm significantly improves coverage over random selection but only slightly compared to the naive approach. If there existed more coverage overlap between participants, the performance gap between greedy and naive algorithms would widen. This indicates that our recruitment framework is more

useful when campaigns have a limited geographic scope (neighborhoods, city blocks) and have participants with higher mutual coverage.

Several individuals participated in multiple campaigns. When participant performance, in terms of sampling likelihood, was compared across campaigns, the individuals on the extremes, either on the high end where their reputation was above normal or vice versa, generally remained at those levels (top or bottom 5 in one campaign stayed in that same range in the others). This indicates that there is potential in using previous performance in similar campaigns to bootstrap reputation models. But a larger study with more varied participants needs to be done to verify this conclusion. Also, in our campaign set, participants grew tired of collecting samples if the campaign lasted for an extended period of time. When individuals performed the campaigns for an additional week, their reputation was much lower. This points to the usefulness of the progress review step to check up on participants especially in long running campaigns.

7.2 Participant Experience Feedback

Participants were asked to fill out post-campaign surveys on their experience in performing the data collections. In terms of capturing data on the mobile phones, participants indicated that it was important that the act of data capture should be streamlined so that it can be repeated rapidly. Many participants also wanted mobile visualizations to help them participate more effectively. For instance, individuals desired a map interface colored by campaign coverage needs and an augmented reality browser to help discover nearby locations for participation. When asked if they would change their routines to participate in campaigns, most indicated that they would be willing to adhere to minor diversions but drastic changes would require extra incentives. Finally, participants stated that daily contribution summaries and in situ reminders would help increase participation.

7.3 Future Work

There are many opportunities to enhance the recruitment framework. The current system relies solely on past coverage and participation behavior. But contributors might be aware of impending changes in their schedule or habits. Individuals could specify a level of service they are willing to offer, and organizers could weight this projection based on participants' profiles and past negotiation fulfillments. Another area of exploration is whether more complex incentive models can help fix sensing gaps caused by inconsistent participants. Bonuses can be given if participants fill immediate campaign needs, and incentives can be scaled depending on context. Finally, the recruitment system should explicitly consider participant sampling bias. Ground truth from independent sources and parameters learned from all participant submissions can quantify this behavior.

Acknowledgments. This work is supported in part by NSF Cooperative Agreement #CCR-0120778 and NSF Grant #CNS-0627084. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding entities.

References

1. Campbell, A., Eisenman, S., Lane, N., Miluzzo, E., Peterson, R.: People-centric urban sensing. In: Proceedings of WiCÓM, pp. 18–32. IEEE, Los Alamitos (2006)
2. Burke, J., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., Srivastava, M.: Participatory sensing. In: Proceedings of WSW, pp. 1–5. ACM, New York (2006)
3. Paulos, E., Honicky, R., Hooker, B.: Citizen Science: Enabling Participatory Urbanism. In: Urban Informatics: The Practice and Promise of the Real-time City (2008)
4. Shallwani, S., Mohammed, S.: Community-based participatory research: Training Manual for Community-Based Researchers. Aga Khan University (2007)
5. Cooper, C., Dickinson, J., Phillips, T., Bonney, R.: Citizen science as a tool for conservation in residential ecosystems. *Ecology and Society* 12(2), 11 (2007)
6. Lu, H., Lane, N., Eisenman, S., Campbell, A.: Bubble-sensing: Binding sensing tasks to the physical world. In: Pervasive and Mobile Computing (2009)
7. Gaonkar, S., Li, J., Choudhury, R., Cox, L., Schmidt, A.: Micro-blog: Sharing and querying content through mobile phones and social participation. In: Proceedings of Mobicys, pp. 174–186. ACM, New York (2008)
8. Reddy, S., Shilton, K., Burke, J., Estrin, D., Hansen, M., Srivastava, M.: Using Context Annotated Mobility Profiles to Recruit Data Collectors in Participatory Sensing. In: Choudhury, T., Quigley, A., Strang, T., Suginuma, K. (eds.) LoCA 2009. LNCS, vol. 5561, pp. 52–69. Springer, Heidelberg (2009)
9. Paxton, M., Benford, S.: Experiences of participatory sensing in the wild. In: Proceedings of Ubicomp, pp. 265–274. ACM, New York (2009)
10. Shilton, K.: Four billion little brothers?: Privacy, mobile phones, and ubiquitous data collection. *Communications of the ACM* 52(11), 48–53 (2009)
11. Hong, J., Landay, J.: An architecture for privacy-sensitive ubiquitous computing. In: Proceedings of Mobicys, pp. 177–189. ACM, New York (2004)
12. Ashbrook, D., Starner, T.: Using GPS to learn significant locations and predict movement across users. In: Personal and Ubiquitous Computing, pp. 275–286 (2003)
13. Kim, M., Kotz, D., Kim, S.: Extracting a mobility model from real user traces. In: Proceedings of Infocom, pp. 1–13. IEEE, Los Alamitos (2006)
14. Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., Terveen, L.: Discovering personal gazetteers: an interactive clustering approach. In: Proceedings of GIS, pp. 266–273. ACM, New York (2004)
15. Liao, L., Fox, D., Kautz, H.: Location-based activity recognition using relational Markov networks. In: Proceedings of IJCAI. AAAI, Menlo Park (2005)
16. Bhattacharya, A., Das, S.: LeZi-update: an information-theoretic approach to track mobile users in PCS networks. In: Proceedings of Mobicom, pp. 1–12. ACM, New York (1999)
17. Soh, W., Kim, H.: Dynamic guard bandwidth scheme for wireless broadband networks. In: Proceedings of Infocom, pp. 572–581. IEEE, Los Alamitos (2001)
18. Hariharan, R., Toyama, K.: Project Lachesis: parsing and modeling location histories. In: Egenhofer, M.J., Freksa, C., Miller, H.J. (eds.) GIScience 2004. LNCS, vol. 3234, pp. 106–124. Springer, Heidelberg (2004)
19. Krumm, J., Horvitz, E.: Predestination: Inferring destinations from partial trajectories. In: Proceedings of Ubicomp, pp. 243–260. ACM, New York (2006)
20. Camp, T., Boleng, J., Davies, V.: A survey of mobility models for ad hoc network research. *Wireless Communications and Mobile Computing* 2(5), 483–502 (2002)
21. Lee, K., Hong, S., Kim, S., Rhee, I., Chong, S.: SLAW: A Mobility Model for Human Walks. In: Proceedings of Infocom, pp. 855–863. IEEE, Los Alamitos (2009)
22. Hsu, W., Dutta, D., Helmy, A.: CSI: A Paradigm for Behavior-oriented Delivery Services in Mobile Human Networks. *ACM Transactions on Networking* (2008)

23. Ghosh, J., Beal, M., Ngo, H., Qiao, C.: On profiling mobility and predicting locations of wireless users. In: Proceedings of REALMAN, pp. 55–62. IEEE, Los Alamitos (2006)
24. Eagle, N., Pentland, A.: Reality mining: sensing complex social systems. Personal and Ubiquitous Computing 10(4), 255–268 (2006)
25. Schlosser, A., Voss, M., Bruckner, L.: Comparing and evaluating metrics for reputation systems by simulation. In: Proceedings of Trust in Agent Societies (2004)
26. eBay: The worlds marketplace, <http://ebay.com>
27. Amazon: Online shopping center, <http://amazon.com>
28. Jøsang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. Decision Support Systems 43(2), 618–644 (2007)
29. SlashDot: News for nerds, <http://slashdot.com>
30. Ganeriwal, S., Balzano, L., Srivastava, M.: Reputation-based framework for high integrity sensor networks. ACM Transactions on Sensor Networks 4(3), 15 (2008)
31. Jøsang, A., Ismail, R.: Beta reputation system. In: Bled eConference, pp. 324–337 (2002)
32. Howe, J.: The rise of crowdsourcing. Wired Magazine 14(6) (2006)
33. Amazon: Amazon mechanical turk, <http://mturk.com>
34. GURU.com: Freelancer network, <http://guru.com>
35. Slijepcevic, S., Potkonjak, M.: Power efficient organization of wireless sensor networks. In: Proceedings of ICC, pp. 472–476. IEEE, Los Alamitos (2001)
36. Kansal, A., Kaiser, W., Pottie, G., Srivastava, M., Sukhatme, G.: Reconfiguration methods for mobile sensor networks. ACM Transactions on Sensor Networks (2007)
37. Kapadia, A., Triandopoulos, N., Cornelius, C., Peebles, D., Kotz, D.: AnonySense: Opportunistic and privacy-preserving context collection. In: Indulska, J., Patterson, D.J., Rodden, T., Ott, M. (eds.) PERVASIVE 2008. LNCS, vol. 5013, pp. 280–297. Springer, Heidelberg (2008)
38. Krause, A., Horvitz, E., Kansal, A., Zhao, F.: Toward Community Sensing. In: Proceedings of IPSN, pp. 481–492. ACM, New York (2008)
39. Zheng, Y., Liu, L., Wang, L., Xie, X.: Understanding transportation mode based on GPS data for Web applications. ACM Transactions on the Web (2009)
40. Mun, M., Estrin, D., Burke, J., Hansen, M.: Parsimonious Mobility Classification using GSM and WiFi Traces. In: Proceedings of EmNets. IEEE, Los Alamitos (2008)
41. Reddy, S., Burke, J., Estrin, D., Hansen, M., Srivastava, M.: Determining transportation mode on mobile phones. In: Proceedings of ISWC, pp. 25–28. IEEE, Los Alamitos (2008)
42. Everytrail: Geotagging with everytrail, <http://everytrail.com>
43. FourSquare: Explore your city, <http://playfoursquare.com>
44. Kang, J., Welbourne, W., Stewart, B., Borriello, G.: Extracting places from traces of locations. Mobile Computing and Communications Review 9(3), 58–68 (2005)
45. Froehlich, J., Krumm, J.: Route prediction from trip observations. In: SAE (2008)
46. Khuller, S., Moss, A., Naor, J.: The budgeted maximum coverage problem. Information Processing Letters 70(1), 39–45 (1999)
47. Nemhauser, G., Wolsey, L., Fisher, M.: An analysis of approximations for maximizing submodular set functions. Mathematical Programming 14(1), 265–294 (1978)
48. Eagle, N., Pentland, A.: Eigenbehaviors: Identifying structure in routine. Behavioral Ecology and Sociobiology 63(7), 1057–1066 (2009)
49. Gonzalez, M., Hidalgo, C., Barabasi, A.: Understanding Individual Human Mobility Patterns. Nature 453(7196), 779–782 (2008)
50. Teacy, W., Patel, J., Jennings, N., Luck, M.: Coping with inaccurate reputation sources. In: Proceedings of AAMAS, pp. 997–1004. ACM, New York (2005)

Out of the Lab and into the Fray: Towards Modeling Emotion in Everyday Life

Jennifer Healey¹, Lama Nachman¹, Sushmita Subramanian¹, Junaith Shahabdeen¹,
and Margaret Morris²

¹ Future Technology Research, Intel Labs, Santa Clara, CA

² Digital Health Group, Intel Corp., Portland, OR

Abstract. We conducted a 19 participant study using a system comprised of wireless galvanic skin response (GSR), heart rate (HR), activity sensors and a mobile phone for aggregating sensor data and enabling affect logging by the user. Each participant wore the sensors daily for five days, generating approximately 900 hours of continuous data. We found that analysis of emotional events was highly dependent on correct windowing and report results on synthesized windows around annotated events. Where raters agreed on the timing and quality of the emotion we were able to recognize 85% of the high and low energy emotions and 70% of the positive and negative emotions. We also gained many insights regarding participant's perception of their emotional state and the complexity of emotion in real life.

Keywords: Affective computing, emotional sensing, mood detection.

1 Introduction

Today's mobile devices allow far more than making phone calls and browsing the web. Thanks to advances in sensing, higher computation power and continuous connectivity, many new applications are emerging from logging physical activity, to measuring and communicating individuals' vital signs, to locating nearby services and friends. Due to their proximity to the users throughout their day, these devices provide a continuous and comprehensive perspective of the user. In our research, we build upon this accessibility aspect to monitor people's emotional state throughout the day. This can be an extremely effective tool for self reflection and self help, especially when coupled with the detection of other contexts, such as activity and social interaction. Awareness of one's current emotional state is a necessary step in the ability to reflect on one's emotional patterns across time and situations. This self-reflective ability, sometimes called mindfulness, is associated with both physical and mental health [1]. A variety of clinical and self-help programs for stress reduction revolve around mindfulness training [2].

Emotion and its physiological correlates have been rigorously studied by psychophysologists. Most psycho-physiological experiments are conducted in a laboratory environment where emotional responses are either performed or primed. This laboratory research reduces the ambiguity of the ground truth determination and focuses on the emotional recognition. However, these laboratory measurements may not reflect

the ranges and patterns of emotional experiences that are present in everyday settings. Our intent was to focus precisely on the complex, noisy emotions that emerge in everyday life. To this end, we conducted an experiment on 19 users for 5 days each, in which we monitored heart-rate data, Galvanic Skin Response data, and physical activity through accelerometer data. Participants were instructed to log their emotional state on smart phones. This self report data was associated with the sensor data and used to develop models for passive monitoring of emotional states. In this paper, we describe our system design for sensor and annotation collection, our experiment design and our data analysis. We also present challenges we encountered in establishing ground truth and their implications for future research design.

2 Related Work

A long history of research has examined the physiology of emotion. Emotion theorist William James first began correlating physiological responses to emotion 1884[3]. Karl Jung used GSR fluctuations to identify “negative complexes” in word association tests in 1906 [4] and the first lie detectors, where changes in GSR and HR were related to guilty stress, were introduced in the 1940s [5]. Recent work in affective computing [6] has for the most part also involved laboratory situations. The majority of reported recognition rates are gathered through priming by stimuli or asking participants to perform an emotion, each of which can cause non-emotion based physiological change. There are many valid reasons for these controlled experiments: the monitoring equipment was traditionally large and difficult to move, real emotions are often complex and difficult to reliably stimulate and in the real world are often caused by events that would be considered too cruel to cause intentionally. Some experiments have ventured into the real world, but were still very constrained and used priming. For example, Picard’s 2005 study measured driver’s stress reaction in the real world [7], but the stress levels were primed by known driving routes and conditions. These controlled experiments did not focus on capturing the range of emotions present in natural settings.

There have been many instances of capturing emotions in everyday life through emotion journaling. Applied psychologists have often had subjects capture their emotional experience in everyday life by recording them on paper [8]. More recently the logging of experience has been possible on smart phones [9]. These emotion journaling studies have either involved sparse annotations or have primarily been designed for targeted intervention, e.g. purposes of anger or stress management.

Ambulatory physiological recording has been possible for medical purposes since the 1960s with the advent of Holter ECG [10]. Since then various medical telemetry devices have become available, including arm and finger blood pressure, respiration, motion for activity and tremor detection, temperature and galvanic skin response [11,12]. In general, these devices have been clunky, single purpose and designed to measure a specific physical or psychiatric medical condition such as Hypertension, Panic Attacks or Parkinson’s disease. The devices have also mainly been recording devices without significant interaction and where the analysis is done offline by medical professionals.

A new era of mobile sensing is being made possible by the availability of wearable physiological sensors and ultra-mobile computing devices. The combination of these two components into a single system allows real time data recording of physiological signals and real time analysis and interaction [13,14]. New systems are also specifically being designed for robust wearability extreme circumstances, such as the monitoring of children [15]. A recent study, Mobile Heart Health, used wireless ECG and mood sampling to trigger therapeutic interventions on the phone to invite self-awareness and coping in everyday life [16,9].

Our system was designed to automatically monitor physiological responses and correlate these with emotion journaling. We measured both heart rate and galvanic skin response physiological signals. We aimed to capture emotions as they happen in uncontrolled, natural environments, while people are driving, singing, chatting with friends, attending a boring meeting and even while going to the dentist.

3 System Architecture and Design

The system comprises of wearable sensors and an aggregation device. The sensor devices monitor physiological signals, such as heart rate (HR), and galvanic skin response (GSR), along with physical activity. The phone aggregator connects to the GSR platform and Mobile sensing platform (MSP) using Bluetooth, gathers data from these sensors and stores it in a mobile database. The watch (Polar R800) aggregates the data from the polar heart rate sensor using a proprietary radio connection.

3.1 Mobile Sensing Platform

Mobile sensing platform (MSP) [17] was used for monitoring physical activity (see **Fig. 1(a)**). The platform aims at supporting a wide range of applications, like inertial navigation, and user activity inference [18]. The package allows the platform to be worn on the waist (belt clip). MSP is a battery operated device equipped with multiple sensors including a 3D accelerometer, which was used for modeling physical activity. Statistical features like mean, variance, min, and max of all 3-axis of the accelerometer were used to build an adaptive boosting classifier to discern activities like sitting, standing, laying, strolling, brisk walking and running. The accelerometer signal was processed every 5 seconds and the classified decision vector containing the most probable user activity was sent to the aggregator to facilitate analysis of the effect of physical movement on the physiological signal.

3.2 Polar Heart Rate Sensors

Polar WearLink along with RS800 logging watch[19] were leveraged as is for monitoring HR and HRV (see **Fig. 1(b)**). The sensor attaches to a conductive fabric chest belt and transmits data to the RS800 watch where the data is logged. The watch and the logging phone were time synchronized to ensure a common time base across the system. The data from the watch was downloaded using the Polar ProTrainer software [19] software for further analysis.



Fig. 1. Sensor Devices included MSP for activity sensing, Polar HR and SHIMMER GSR

3.3 GSR Sensor

Galvanic Skin Response is a measure of change in the conductivity of the skin due to an individual's psychological state and is widely used as a modality for monitoring stress and mood related changes [20, 15]. The principle of operation of GSR is based on the change in conductance due to the amount of sweat level in the eccrine sweat gland [21]. We are not aware of commercially available GSR solutions that meet our requirements. Hence we developed a sensor board capable of measuring change in conductance and connected it to the SHIMMER platform [22], which acted as the processing and communication unit. The device was harnessed to a wrist band and a neural electrode was attached to the fingers for monitoring the change in conductance (see **Fig. 1(c)**). The data from the sensor board was sampled at 4Hz and transmitted to the aggregator via Bluetooth.

3.4 Mobile Phone Aggregator

An HTC Touch Pro phone was leveraged for data storage and user interface. The phone implemented the software architecture described below and acted as an aggregator for the data transmitted from the sensor devices (MSP, GSR). The data was time-stamped and stored in a mobile database for offline processing. The phone was also used to collect ambient audio data at 11 KHz and stored it into wav files.

3.5 Software Architecture

Fig. 2 describes the software architecture of our aggregator device. It consists of a proprietary framework (Carson Springs) that provides sensor communication, data storage and the ability to plug in application level modules like user prompter, GSR and MSP data processors and user interface. We used the polar heart rate sensor and aggregator as is and the aggregation mechanism is not described here. All the components listed below are implemented in S/W and run on the phone.

Carson Springs Framework: This is an internal framework developed at Intel consisting of four major components, the sensor controller module along with the Bluetooth communication module allowed the application to connect to the sensor nodes to send / receive data. The data exchange module allowed the application level

components to register for data from sensors for processing and connection verification. The MSP data processor module parsed the result vector from MSP and extracted the most likely physical activity. This information was forwarded to user prompter and data storage modules through data exchange. Finally the data storage module comprised of data access and DB Writer acts as an interface to store and retrieve data from the mobile database.

User Prompter/ GSR Analysis Module: The prompter module implemented the annotation reminder logic. User prompting was triggered at thirty minute intervals and when the system detects an interesting signal changes. We developed a naïve processing algorithm for GSR that detected rate of change in the signal and specific patterns in the tonic level to identify an interesting event. The information from MSP was used to filter out events generated during active states. The events generated during sedentary state were used to trigger an annotation prompt by playing a sound file on the phone. In order to minimize annoyance to the user, we programmed the algorithm to prompt the user at most once every 15 minutes. Events that occurred within 15 minutes of a previous event were not prompted. The signal events and the periodic events were stored in the database along with the annotations to facilitate further processing.

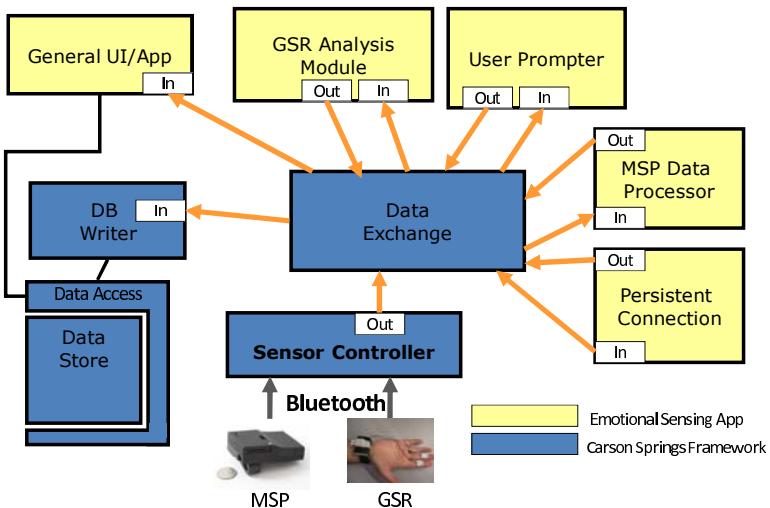


Fig. 2. Sensing and aggregation architecture

Persistent Connection: The distributed nature of the system introduces several opportunities for connection loss between the aggregator and sensor devices. The persistent connection module monitors the sensor data traffic to determine connectivity with the sensor devices. If the data flow is broken from a sensor device the module identifies the disconnected sensor device and periodically attempts to reconnect. It also communicated the loss of connection to the UI. Connection loss is a common problem in wireless body worn devices as discussed in [23]. The ability to reconnect significantly improved the reliability of data collection.

User Interface: The main purpose of the module was to allow the user to start/stop the data collection and annotate their emotional states. It also provided feedback to the user about sensor connectivity. The annotation part of the interface is described in detail in section 4.2.

Time Synchronization: Use of off the shelf polar heart rate monitor prevented us from having a single aggregator due to radio incompatibility. We had to synchronize both phone and watch aggregator to a specific laptop to ensure synchronization of heart rate data with data from other sensors. The laptop was in turn synchronized to an NTP server through the Intel network. The synchronization was repeated daily during the download/interview process to compensate for clock drift in the platforms.

4 Experiment and Study Design

Our main goal for the study was to gather “ground truth data” by having participants report their affective states for a period of days. Alongside these self-reports, we used sensors to record physiological signals and audio of the participant. The ground truth data was intended to help us develop inference algorithms for affective state detection in ambulatory settings and to understand what is possible to detect via sensing.

4.1 Recruitment

Nineteen full-time professionals enrolled to participate in the study (12 men and 7 women). Our participants were a convenience sample of colleagues at Intel Corporation. These full time professionals were predominantly engineers ($n=16$) and the rest worked in marketing or management. No participants were on heart-altering medication. The majority of our participants were in their late 20s and 30s, and 6 were older than 35. Participants were recruited via email sent to a pool of our contacts and referrals.

4.2 Study Protocol

Participation involved an introduction meeting, daily interviews, and a final interview. In the introductory meeting, participants reviewed a consent form, and the process for annotating their moods and operating/wearing the sensor. Daily interviews, conducted at the end of each work day, were held to understand participants’ annotations and to add annotations that they did not make during the day. These interviews began with guided open-ended questions about participants’ affective states during the day, and included queries about high and low points in their day, comparison of the morning and afternoon time segments, and comparisons of that day to the previous one. Next we asked targeted questions about specific times of the day based on their sensor data and annotations. Lastly we reviewed the day’s sensor data with our participants. The final interview included a review of the entire week and a discussion of their high/low points of the week and any insights participants gleaned about their emotional patterns. We also used this interview to gather feedback on the trial, such as wearability of the devices and/or usability of the interface. For the duration of the study, participants wore the three sensors described (GSR, heart rate, and accelerometer) and

carried an HTC Touch Pro smart phone for eight or more hours a day. They were instructed to log their affective states on these phones every time they experienced a change in their affective state or when their behavior might influence their physiological data, e.g. eating, drinking of caffeinated beverages, or adjusting the electrodes). In total, participants annotated anywhere from 5 to 40 times a day, averaging 19 annotations a day.

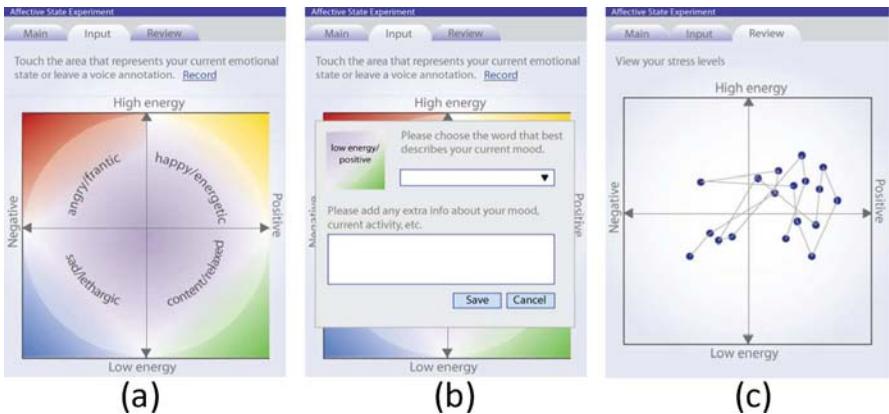


Fig. 3. Users annotate their emotions using the Mood Map (a) and emotion term specification and free text description (b). They also are able to immediately view their graph annotation entries for the day (c).

For each annotation, participants indicated their current affective state as a point on the Mood Map [16,9]. The Mood Map was previously developed as a touch screen translation of the circumplex model of emotion [24] that allowed for intuitive and accurate mood reporting. This interface was found to invite self-awareness and reflection. The circumplex model is well evidenced in psychology research and the Mood Map was extensively tested and revised in previous field tests [9]. In this 2-dimensional map, the Y axis represents low to high arousal and the X axis represents negative to positive valence. As a result of iterative testing, a couple of revisions were made to the interface: the arousal axis was labeled as “energy” and emotional descriptors (e.g., happy, excited, angry) were removed from the quadrants. These terms were intended as a loose guide for Mood Map entries, but gave some participants the incorrect impression that they needed to rate the valence and intensity of each term. For the current study we used the Mood Map without emotional terms in the quadrants, but added mood word selection as a second stage of input. This two stage entry allowed the Mood Map coordinates to be collected independently from the word selection (See Fig 3(b)). A set of affective state labels were chosen based on the words used in early testing of the Mood Map [16] and other terms commonly used by our participants in pilot studies. An option for “other” allowed participants to specify a word not in the menu. We also included an area for freeform text which was intended both for data patterning/validation and as stimulus for daily interviews.

Participants could also review the points they had selected throughout the day on the Mood Map as shown in Fig. 3(c). Sensor data was not displayed during the day to avoid influencing participants' behavior. However, participants could review their sensor data at the end of each daily interview. The sensor data, like the annotations, were used as stimuli for interview discussion. The peaks and valleys were used to trigger discussion about emotional experiences that may have been forgotten.

In addition to these annotations, the smart phones allowed participants to capture audio recordings. Again, these recordings were used to aid recall in daily interviews. We also requested participants' permission to have an automated system analyze these audio recordings to extract auditory features, such as pitch and volume, without processing their speech content. Participants could opt-in to this part of the study and could control when they wanted to capture these audio recordings.

4.3 Incentives

We wanted to recognize participants' time and efforts in this trial for carrying four extra devices, making frequent annotations, and making room in their schedules for daily interviews and troubleshooting. To alleviate these burdens and to encourage active engagement in the study, we used an approach of compensating participation with a base structure (an Apple iPod shuffle) and incremental rewards; specifically iTunes gift cards ranging from \$5-\$20 per day based on the number of annotations they made. An annotation was considered as a mood map selection, a mood word selection, and extra information that the participant entered about their context at the time. We gave \$5 for up to 10 annotations/day, \$10 for up to 20 annotations/day, \$15 for up to 30 annotations/day, and \$20 for over 30 annotations. We also awarded a bonus gift card each week to the participant who made the most annotations.

5 Data Analysis and Key Learnings

Our initial approach to the data analysis was to assume that users would annotate emotional events soon after experiencing them. The system design included software algorithms that automatically detect physiological events as the users experience them and prompt them to annotate. These algorithms were derived from previous experiments in emotion recognition and long term stress detection. In our initial analysis plan, we allowed for a one minute "eye-closing" period immediately preceding each emotional event annotation. During this eye-closing period we did not "look" at the data because we assumed the emotional response would be corrupted during this time due to the reflection inherent in the act of annotation. Therefore we only used the data preceding this period for analysis. We experimented with fixed time windows of different lengths as shown in Fig. 4. The signal during the eye-closing period is highlighted in red and each of the preceding windows highlighted in a different color. For example, the five minute window would include data from the blue, pink, black and green segments as indicated by the line labeled "5 min" extending across this period.

From each of these fixed windows we planned to extract features of the GSR and heart rate that have been previously hypothesized to differentiate emotions [25,26,20] These included: the mean and variance; median and inter-quartile range as more

robust estimators of average and spread; and features reflecting the overall shape of the signal such as the slope and kurtosis. In addition, we considered features that were specific to each sensing modality, including peak frequency and rise/falls times of the GSR and root mean successive difference (RMSSD) of the heart rate to estimate heart rate variability [9].

From previous studies, we realized that motion would be a confounding factor in the analysis, so we eliminated the data from time periods where the user's physical activity exceeded strolling. This was done using the MSP as mentioned earlier.

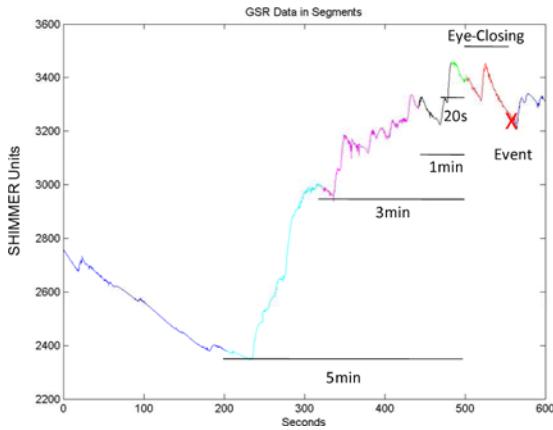


Fig. 4. Fixed time window preceding the user annotation was initially assumed. Sensor data from these time windows was annotated with the emotional event.

For each time window (1 minute, 3 minutes, 5 minutes), sets of features were calculated for both GSR and HR (where valid data existed). Data was labeled according to several aspects based on where the subject had tapped on the Mood Map and the chosen emotion word. For example, to train a “high” vs. “low” arousal classifier, all the segments labeled with a tap in the “high arousal” and “low arousal” section of map were used. Similarly to train the “positive valence” vs. “negative” valence classifier, segments with tap values on the left and right of the mood map were used. We additionally tried to build classifiers based on emotion word clusters. Each of the feature sets was evaluated in WEKA [27] using ten-fold cross validation (every tenth sample is reserved for the test set) and a selection of learning algorithms including: Bayes Net, Naïve Bayes, Adaboost, and the J48 Decision Tree. Results were analyzed for all subjects collectively and for each subject individually. The results showed that the only classifier to perform better than naively guessing the most popular class was the J48 decision tree for an individual, unfortunately these trees proved to be over-fit. We tried different methods of dividing the training and test set, but balanced sets of “high” vs. “low” arousal features were still showing 51% error rates. Finally we included all of the data in the training set, and even when testing on data the classifier

was trained on, the error rate for the Bayes Net classifier was still 50%. This convinced us that this fixed window data did not contain differentiating information and could not be used to develop a classifier.

We discovered that the data features were highly dependent on both window length and placement. To illustrate this point, Fig. 5 shows the GSR signal of the same event viewed through three different time windows. Features extracted from each of these windows vary considerably as demonstrated in Table 1. As a result, choosing the correct time window is crucial.

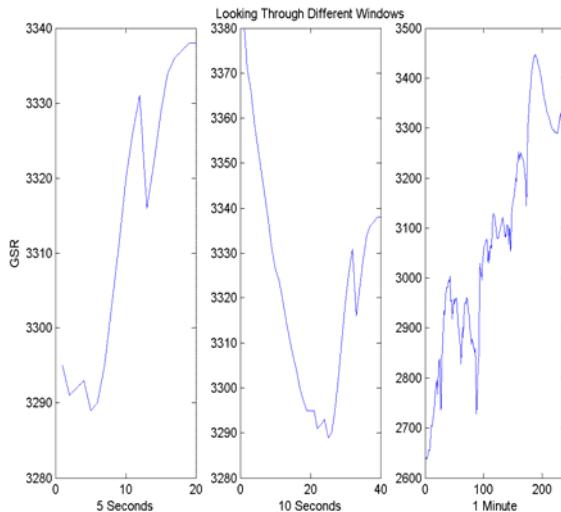


Fig. 5. Effect of time windows: The different figures shows the same GSR signal mapping of the same event using 3 different time windows

Table 1. Effect of time window selection on feature calculation

Feature	5 sec window	10 sec window	60 sec window
Mean	3314	3312	3083
St Dev	18.5	23	217
Slope	43	-69	697

We also discovered that users didn't necessarily annotate directly following an emotional event resulting in variable time delay between the start of an emotional event and the annotation. For example, one user commented in their free text notes: "annoyed in an argument 15 minutes ago". It was also evident from a review of the annotations that different emotional episodes lasted different lengths of time. For example one participant reported eight emotions in the space of an hour with no two successive emotions in the same affective quadrant. Another user reported being either stressed irritated or frustrated over a single cause for almost three hours.

From these observations we determined that having a time window customized to each emotional experience was important for extracting the correct features for analysis. Since these windows were not available from the subjects directly, we used all available evidence (emotion words, taps and end of day interviews) to make the best estimate of these windows. An initial rater (R1) looked to group extended periods of similar emotions periods (e.g. angry, irritated, and stressed) into longer windows called “emotion arcs.” Each arc was then assigned a valence and an energy label where the valence was labeled as positive, neutral or negative and the energy was labeled as high, neutral or low. R1 used the GSR signal as a guide to determine likely transitions and to determine where the data was valid. Features were extracted from 80% of the R1 arcs and used to train two BayesNet classifiers using WEKA [27], one for energy and one for valence. The remaining 20% of the R1 arcs were withheld from the classifier and given to a second rater, R2, who independently assessed the valence and energy and was allowed to adjust arc duration for the test set.

R2 agreed on the timing for 42% of the arcs (tBoth). In most cases when R2 disagreed on the time it was because R2 saw a transition (e.g. from high to neutral) and ended the arc sooner. We looked at how well the raters agreed with respect to energy and valence over the agreed arcs (tBoth) and over all of time periods chosen by R2 (tR2). The results were similar for both sets of windows. **Table 2** shows that the raters agreed exactly on one of the three energy levels (high, neutral, low) for 50% of the tBoth windows (46% for tR2) and agreed exactly on the valence level (negative, neutral, positive) for 44% of the tBoth windows (64% for tR2). The disagreements were rarely in opposition and raters were within one emotion level of each other (e.g. “high” vs. “neutral”) 81% of the time for energy and 93% of the time for valence over the tBoth windows with similar results for tR2.

Table 2. Agreement between emotion arc ratings

	Exact tBoth	+/- 1 tBoth	Exact tR2	+/- 1 tR2
Energy	50%	81%	46%	82%
Valence	44%	93%	64%	91%

We created four test sets from the 20% of the data withheld from the classifier, two sets using features from data extracted from the tR2 windows and two sets from the tBoth windows. The two sets were sets where R1 and R2 agreed exactly on the emotional state, likely indicating obvious expression of the emotion, and where R1 and R2 disagreed on the emotional state, likely indicating a more ambiguous expression of emotion. The results are shown in **Table 3**.

Table 3. Recognition accuracy using emotion arcs

	Disagree (tR2)	Disagree (tBoth)	Agree (tR2)	Agree (tBoth)
Energy accuracy	55%	33%	80%	85%
Valence accuracy	50%	54%	60%	70%

These results show that the highest recognition accuracy was obtained when raters agreed on time windows and emotion labels. These instances were likely more prototypical expressions and therefore easier to discriminate. Analysis of the energy classifier showed that the most differentiating features were GSR mean and the slope. Analysis of the valence classifier showed that most differentiating features for valence were the GSR mean, the maximum peak rise time of the orienting response and the maximum slope.

Previous results have reported in lab discrimination of 66-92% for four quadrant arousal valence discrimination in the lab [6] and 78-86% in intelligent tutoring systems for high vs. low discrimination of Confident, Frustrated, Excited and Interested [28]. If we consider only the least ambiguous emotion states in our test set, our results of 85% for high and low arousal and 70% for positive and negative valence approach these results. However, in the real world we face the problem of ambiguous emotional states which may confound real time discrimination using physiology alone. This problem may be solved by modeling the user's context. Carroll and Russell showed that context was a key element in human emotion discrimination. Using only prototypical facial expressions, human recognition accuracy was 69%, but with supporting context information recognition increased to 74-100%. For our system, supportive context information might be added by modeling what the user is doing and who they are with as well as by incorporating other sensor channels such as voice analysis and facial expression which have been shown to increase recognition accuracies [28]. Given the current low overall accuracies, the best use for the current system may be using the results in aggregate over longer periods of time, for example comparing afternoons where the user went to lunch with friends versus eating at his desk over several months. In aggregate, the system should be able to differentiate between these two cases even if the instance by instance accuracies are low. These long term results could give the user insight into the real effects of daily choices and aid in long term behavior planning for better life balance.

6 Discussion

6.1 Difficulties of Accurate Self Reporting

Capturing truly objective “ground truth” data about people’s affective states was challenging due to apparent disparities between the Mood Map points, affective state words, and participants’ descriptions in the freeform text and interviews.

We compared the specified affective words with Mood Map coordinates, finding a wide range of points on the Mood Map across participants and even across an individual’s annotations. An example below (Fig. 6(a)) illustrates how the word “calm” correlated with points that were in both the low and high energy quadrants of the map. And, though most of the points were in the positive half of the graph, there were a handful of points in the negative half of the graph. The spread was surprising and we consider several explanations.

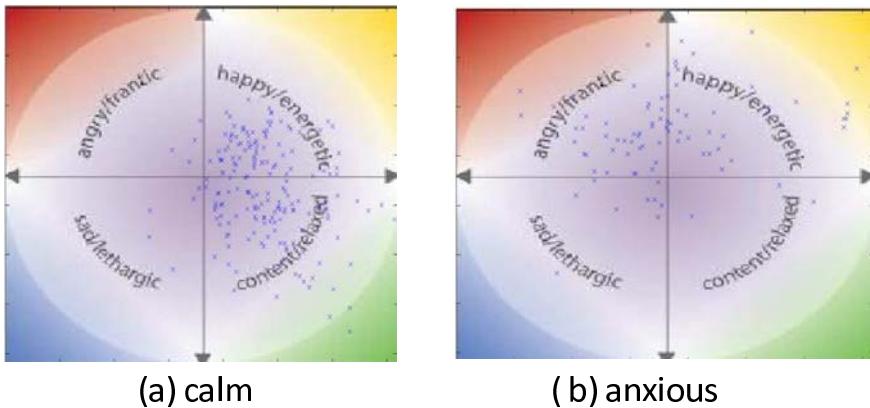


Fig. 6. Selection of “calm” were associated with both high and low energy. Selection of “anxious” were associated with both positive and negative valence. These ranges could reflect a misinterpretation of the center of the Mood Map or the complexity of an emotional state.

Different responses to the Mood Map in the current study may relate to the different goals and durations of the studies. In contrast to the previous Mood Map work, the current study focused on identifying specific emotional states for use in a machine learning model. And in contrast to the month long deployment of the past study, the current study gathered data over a one week period. In the longer study, participants calibrated their responses over time [33]. In the current study, some participants appeared to interpret the center of the graph as (0,0) and therefore the lowest in terms of energy. This tendency would explain why many of the annotation points were collected at the center of the graph rather than towards the bottom of the graph, even when participant explicitly described themselves as low energy. Alternatively, it is possible that people felt energetically calm in some moments and sleepily calm in others. And while calm is usually associated with positive experiences, it can also be associated with boredom, a negative state. This exemplifies the complexity of emotion and emotional measurement.

Another potential reason for the wide distribution of annotation points is that affective states are complex and generally one word does not summarize a state in some consistent way. Participants chose the “best fit” label and sometimes their understanding of a label was broad enough to be applied to a few different emotional states. For example, the word “calm” was used to describe times when a person was meditating (actively trying to achieve a state of positive peacefulness), but then also anytime a person was in a “neutral” or “fine” state. Similarly, one commonly reported state was “happy”, for instance “happy to be out of that meeting” or “happy was stressed, but happy now that this problem is resolved”. Happy in these contexts was describing a sense of relief, which is a very different than the sense of happiness when “eating cookies!”. The word “anxious” was another label that we found had surprising variance across on the Mood Map, spanning both the positive and negative quadrants (Fig. 6 (b)). We found that the word “anxious” could be correlated both with

hopefulness and stress/nervousness. This finding was observed in previous research on the Mood Map. [9]

Another issue that we came across in our study was a strong trend towards positive affective states in terms of Mood Map annotations. **Fig. 7** shows the mean for the collective values of all the taps associated with each of the emotion words. Most emotions words show means trending towards the positive side compared to our initial assessment of the location of such words. There are a couple possible reasons for this positive bias in the data. One possibility is people's desire to be perceived as positive, which would significantly affect their annotations. People seemed to view annotations that were meant to describe a specific moment in their day as a reflection on their overall self. For example, some picked a label such as "annoyed" and during the interview they would make it clear how irritated they were, and yet their quadrant point would be somewhere in the right half of the graph on the positive side. This positive bias was prevalent across participants in their graph annotations. This bias may have been more evident in the graph vs. the labels because the graph axis was explicitly labeled "positive" and "negative". Also, the review pane of the interface allowed participants to review only their graph clicks. Both of these factors may have influenced people to want to annotate a general positive state with which they wanted to represent themselves. One participant explained this by stating "There were several times when I picked a mood word for a specific annotation, and 'on purpose' placed my mood in a quadrant that might have seemed contradictory to my mood word. This wasn't because I was uncomfortable reporting my information, but rather that I perceived a difference between a specific "in the moment mood" as indicated by a word, and my overall general mood." "In general, I am a person who spends most of my day in a positive mood, but there are incidents throughout the day that can annoy, frustrate, etc.. If I was annoyed/frustrated, I would denote that point, and it wouldn't have been uncommon for me to sometimes list that I was still in a positive mood."

Retrospective bias may also be at work. People often remembered an event very differently after the event occurred rather than during [31,32]. For example, one of our participants told us that he was very nervous for a presentation he was planning to give. In the days leading up to the presentation he expressed concern and anxiousness about the upcoming presentation and he described that he was very stressed beforehand because his manager asked him to change several things shortly before the presentation was scheduled to begin. However his annotations (made immediately after he presented) stated that he was calm and positive. He explained this saying "Yeah, I was stressed before that presentation, but I was fine. It wasn't really a negative feeling." This discrepancy appears to reflect coloring of his past mood by his current mood, a finding well established in cognitive psychology. As mentioned, we designed the system to prompt the user during key moments when we detected interesting sensor data activity to encourage them to annotate and explain these key moments. However, almost all of our participants turned off the sound or vibrations on the phone during the study. Since the phone was with them at all times, they did not want the phone to accidentally ring during a meeting, so they did not hear or receive any of these prompts. Also, although our study was designed to have people annotate as often as possible while "in the moment", sometimes this was not possible such as during a dentist visit or while giving a presentation.

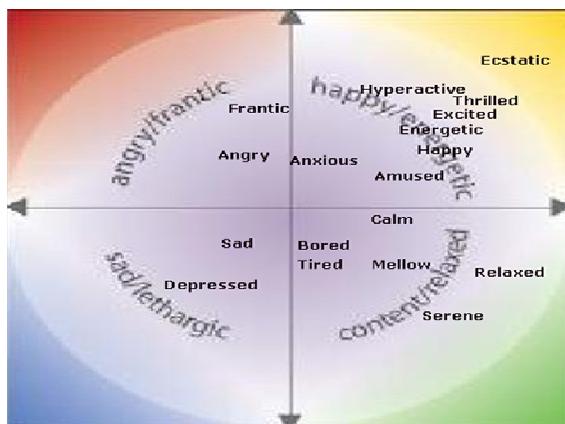


Fig. 7. Each emotion word is plotted based on the mean of all taps associated with each word

6.2 Effective Study Design Decisions

We gained insight from this study into a few design decisions that had significant positive impact on gathering ground truth data. We learned that the process of reviewing the data with participants on a daily basis motivated them to make more frequent and better targeted annotations. We found that the data really engaged participants and often incentivized them to want to take more detailed notes about their happenings once they saw that their sensor data captured finer grained details than they expected.

We also found that having a step ladder of gifts (daily iTunes gift cards) in addition to a base participation gift (iPod shuffle) was an effective means to maintain motivation throughout the week. There are several approaches we could have taken towards rewarding participants. We selected this approach, believing that it would bias participants towards responding, but not responding in a particular direction. A potential side effect of this approach is that participants may exaggerate what constitutes for a change in state. However, we did not reward based on the content of the annotation, but rather on the amount of information within an annotation. Additionally, winners of the bonus gift card were excited by their accomplishment.

Conducting the interviews at the end of each work day was crucial for effective participant reflections. On a few occasions we had to postpone these end-of-day interviews to the morning after and we found that participants had a great deal of trouble remembering events from the previous day. Our interviews were most successful when we did these reflections at the end of the day while the events of the day were still fresh in people's minds.

Finally, of the separation of the graph and the mood words was illuminating. We found that participants had very different notions of which words associated with an area on the graph and this association constrained their input. In the first week of the trial, we tested loading specific mood words in our dropdown depending on which quadrant the participant selected. Participants described that they sometimes chose a point on the Mood Map that they did not feel described their state in terms of valence

and energy levels just so they could select a specific mood word. Although separating the graph and mood words brought up a lot of the inconsistency issues as mentioned earlier, it also allowed participants to describe their state in a more accurate way.

6.3 System Level Issues

The system described above is slightly different from our original design and the changes mainly aimed at ensuring a reliable data collection. The major change is the heart rate sensor, our initial design consisted of the SHIMMER [22] device with an ECG sensor that connected to the phone via Bluetooth. Limited availability of the SHIMMER-ECG device, coupled with the hardware failures during the first two weeks of the experiment prevented us to continue the usage of SHIMMER so we switched to Polar instead. This change also affected the user prompter module in the aggregator, where the logic for triggering events based HR changes was unused due to the lack of data.

6.4 Data Quality Issues

Gathering physiological data from novice users through a distributed wireless system is challenging due to many factors. For this study these included: subjects wearing the sensors incorrectly, subjects failing to fully charge the sensors, subjects re-using the pre-gelled electrodes after the conductive gel pad had been compromised, chest straps losing contact with the body due to movement and lack of moisture, subjects losing wireless connectivity by walking away from the device and subjects accidentally turning off either the sensors or the phone. As a result of the above issues our data yield was less than 50% of the data that could have been harvested from both GSR and HR. In near future experiments, we plan to alleviate these issues with more detailed documentation and videos describing the system and how to wear the sensors, asking the subjects to thoroughly wet the HR strap and instructing subjects to use fresh GSR pads every time the GSR becomes dislodged. Additionally, we plan to have automatic data quality assessment algorithms running on the system, have data replicated back to the server periodically for visual inspection and make the system more robust to disconnects. Our vision for the future would be that these sensors would ultimately disappear into clothing and that HR and GSR could be sensed through fabric electrodes on the body [30], near the chest for HR; and in socks or insoles to measure GSR from the foot [13]. These sensors would be ultra low power or zero net energy by harvesting motion and heat energy from the body. These devices would have ubiquitous connectivity with any trusted source and data would always find a path back to the server or the user's personal mobile platform.

7 Future Work

In this paper, we have presented the results from a study aimed at capturing and correlating physiological data and emotions. The data collected in this study were intended for the development of inference algorithms for automatic affective state detection in ambulatory settings. We have described a set of key findings and challenges involved in the capture of physiological and emotional data in everyday life, namely issues

with accurate self-reporting of emotion, the varying time spans of emotions, and the fact that energy levels are more easily distinguishable physiologically than valence levels.

In future work we plan to enrich the user annotation experience by allowing customized windowing of events and automatically annotating the user's day with high level activities and people proximity using technologies currently under development. We will address the issues of Mood Map and mood label interpretations, bias, and inconsistencies by focusing on a smaller set of emotions and allowing people to input their state on a spectrum. To better capture other aspects of emotion, we also envision using affective analysis of voice, captured by a mobile device and facial expression analysis when the user is seated in front of a camera-enabled computer. We believe that capturing this data is complementary to the physiological data and such fusion will help improve the valence estimation accuracy. Also, to mitigate privacy issues with voice recording, we plan to extract features from the voice and not record any raw audio. Future work should also address individual differences in emotional reactivity, a complex but important health indicator.

References

1. Langer, E.J.: *Mindfulness*. Perseus Books, USA (1989)
2. Kabat-Zinn, J.: *Coming to Our Senses: Healing Ourselves and the World Through Mindfulness*. Hyperion Books, New York (2005)
3. James, W.: William James writings 1878-1899, chapter on emotion, The Library of America, p. 1992 (1890)
4. Jung, C.G., Montague, D.E.: *Studies in Word Association*. Routledge and K. Paul (1969)
5. Marston, W.M.: *The Lie Detector Test*. R.R. Smith, New York (1938)
6. van den Broek, E., Janssen, J.H., Westerink, J.H.D.M.: Guidelines for Affective Signal Processing (ASP): From Lab to Life. In: Proceedings of the International Conference on Affective Computing and Intelligent Interaction, September 10-12, vol. 1, pp. 217–222. IEEE, Los Alamitos (2009)
7. Healey, J.A., Picard, R.W.: Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems* 6(2), 156–166 (2005)
8. Oatley, K., Duncan, E.: The Experience of Emotion in Everyday Life. *Cognition & Emotion* 8(4), 369–381 (1994)
9. Morris, M., Guilak, F.: Mobile Heart Health: Project Highlights. *IEEE Pervasive Computing* 8(2), 57–61 (2009)
10. Holter, N.J., Gengerelli, J.A.: Remote Recording of Physiological Data by Radio. *Rocky Mountain Medical Journal Colorado Medical Society* 46, 749–752 (1949)
11. Fahrenberg, J., Myrtek, M. (eds.): *Progress in Ambulatory Assessment*. Hogrefe and Huber Publishers (2001)
12. Hofmann, S.G., Barlow, D.H.: Ambulatory psychophysiological monitoring: A potentially useful tool when treating panic relapse. *Cognitive and Behavioral Practice* 3(1), 53–61 (1996)
13. Healey, J.A., Picard, R.W.: Affective Wearables. In: Proceedings of the IEEE 1st International Symposium on Wearable Computers, ISWC, Cambridge, MA USA, October 13-14, pp. 91–97 (1997)

14. Westerink, J., Ouwerkerk, M., de Vries, G., de Waele, S., van den Eerenbeemd, J., van Boven, M.: Emotion measurement platform for daily life situations. In: Proceedings of the International Conference on Affective Computing and Intelligent Interaction, September 10-12, vol. 1, pp. 704–708. IEEE, Los Alamitos (2009)
15. Hedman, E., Poh, M., Wilder-Smith, O., Fletcher, R., Goodwin, M.S., Picard, R.: iCalm: Measuring Electrodermal Activity in Almost Any Setting. In: Proceedings of the International Conference on Affective Computing and Intelligent Interaction, September 10-12, vol. 1, pp. 594–595. IEEE, Los Alamitos (2009)
16. Morris, M.: Technologies for Heart and Mind: New Directions in Embedded Assessment. Intel. Technology Journal 11(1) (2007)
17. MSP Platform description, <http://seattle.intel-research.net/MSP/>
18. Choudhury, T., Consolvo, S., Harrison, B., Hightower, J., LaMarca, A., LeGrand, L., Rahimi, A., Rea, A., Bordello, G., Hemingway, B., Klasnja, P., Koscher, K., Landay, J.A., Lester, J., Wyatt, D., Haehnel, D.: The Mobile Sensing Platform: An Embedded Activity Recognition System. IEEE Pervasive Computing 7(2), 32–41 (2008)
19. Polar USA, <http://www.polarusa.com/us-en/products>
20. Picard, R.W., Vyzas, E., Healey, J.: Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(10), 1175–1191 (2001)
21. Stern, R.M., Ray, W.J., Quigley, K.S.: Psychophysiological Recording, ch. 13, 2nd edn. Oxford University Press, Oxford (2001)
22. SHIMMER: http://shimmer-research.com/wordpress/?page_id=20
23. Wan, C., Sai, P.: Challenges to Building Bluetooth-based Sensing Solutions. In: International Conference on Body Area Networks (April 2009)
24. Russel, J.A., Mehrabian, A.: Evidence for a three-factor theory of emotions. Journal of Research in Personality 11, 273–294 (1977)
25. Levenson, R.W.: Autonomic Nervous System Differences Among Emotions. American Psychological Society 3(1), 23–27 (1992)
26. Ekman, P., Levenson, R.W., Friesen, W.V.: Autonomic Nervous System Activity Distinguishes Among Emotions. Science (221), 1208–1210 (1983)
27. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (1999)
28. Cooper, D.G., Arroyo, I., Park Woolf, B., Muldner, K., Burleson, W., Christopherson, R.: Sensors Model Student Self Concept in the Classroom. In: Houben, G.-J., McCalla, G., Pianesi, F., Zancanaro, M. (eds.) UMAP 2009. LNCS, vol. 5535, pp. 30–41. Springer, Heidelberg (2009)
29. Carroll, J.M., Russell, J.A.: Do facial expressions signal specific emotions? Judging emotion from the face in context. Journal of Personality and Social Psychology 70, 205–218 (1996)
30. Paradiso, R., Loriga, G., Taccini, N.: A wearable health care system based on knitted integrated sensors. IEEE Transactions on Information Technology in Biomedicine 9(3), 337–344 (2005)
31. Blaney, P.H.: Affect and memory: a review. Psychological Bulletin 99, 229–246 (1986)
32. Bower, G.H.: Mood and memory. American Psychologist 36, 129–148 (1981)
33. Morris, M., Kathawala, Q., Leen, T., Gorenstein, E.Q., Guilak, K., Deleeuw, B., Labhard, M.: Mobile therapy and mood sampling: Case study evaluations of a cell phone application for emotional self-awareness (submitted)

The Secret Life of Machines – Boundary Objects in Maintenance, Repair and Overhaul

Matthias Betz

Fraunhofer Institute for Applied Information Technology FIT
53754 Sankt Augustin, Germany

Department of Information Systems and New Media, University of Siegen
57068 Siegen, Germany
matthias.betz@fit.fraunhofer.de

Abstract. The increasing level of automation in tight just-in-time subcontracting relationships in the automotive industry makes the complex, weak structured, knowledge intense and highly cooperative practice of Reactive Maintenance (RM) in Maintenance Repair and Overhaul (MRO) in this branch a demanding and stressful job. In this paper two typical breakdown situations are presented which occurred in a participative observation to gain insights to this field. Based on the analysis of the observations and the existing MRO related IT infrastructure we refer to the theoretical concept of ‘boundary objects’ to understand the practice in this field. Finally, implications for design for a MRO supporting pervasive computing environment are derived from this conceptualization. We highlight the potentials of attaching relevant information to physical objects in place to support and motivate documentation by bridging the physical world of machines with the virtual information space and to enhance the discovering of relevant information in breakdowns situations.

Keywords: maintenance, repair, overhaul, collaboration, boundary-objects, histories, practice, observation, UbiComp, autoID, physical, sensor-networks.

1 Introduction

Modern mass-production, especially the production of automotive-components is complex and time-critical. With the introduction of Lean-Production and the pervasiveness of Just-In-Time logistics the need for storage capacity was reduced at the cost of an increasing dependency on the availability and reliability of production facilities. In production settings with a high level of automation, machines are complex conglomerates of different technologies like pneumatics, hydraulics, electronics, programmable controllers, interfaces to information systems like production data acquisition, enterprise-resource-planning systems etc. Additionally, machines in mass-production are also embedded into complex workflows, i.e. inputs and outputs of different machines are connected in complex interdependencies. Due to the increasing requirements regarding availability and reliability of the technical equipment the routine character of maintenance, repair and overhaul (MRO) work in such settings develops towards a time-critical more “first responder” like work. Beyond that,

the increasing complexity of the production facilities in terms of heterogeneity of machinery-vendors, age, compositions of different equipment types, standards, work-flows, breakdown-situations etc. makes MRO work a hardly predictable, weak structured and knowledge-intense process. To increase the reliability and the availability of the technical equipment in time-critical production settings the support of MRO is an important factor in this domain.

This contribution summarizes the results of an empirical study in a MRO department of a medium-size company for plastic shaping. It focuses on the questions how maintenance workers deal with repair-related information, how this information is related to objects and places and to which extend those objects can be seen as boundary-objects for the involved groups of actors. Consequently – from a HCI perspective – the question arises how maintenance workers could be supported in collaboration, i.e. finding relevant information, artifacts and experts for efficiently bringing facilities back into a productive state. After providing an overview of current concepts and approaches in supporting MRO work with information technology the methods and findings of the empirical work in the MRO department will be summarized. Based on the findings, implications for the design of pervasive IT support of MRO workers will be presented based on the theoretical concept of ‘boundary objects’ (c.f. [1, 2, 3]). Following [3] especially the aspect of information reuse of recorded states of boundary objects in time, i.e. the MRO history of machines, is taken into account.

2 MRO in Research: Terms and Focus in IT Support

Focusing on MRO in the literature there is only little research in the area of maintenance and reliability. Given the significance of MRO to manufacturing competitiveness, it is surprising how little research is being carried out in this area. Within the CSCW research there are some MRO related works, e.g. the investigations on maintenance and repair of copy and printing machines at Xerox. As stated in [4] maintenance is a highly non-routine work with combined social and technical complexities: Individual machines have idiosyncrasies that defy understanding; replacement parts often don't work; work rotation sometimes has one technician invading another's service territory, neither the regular technician nor the interloper knows the history of the machine (which turns out to be very important in this contribution). All these social and technical complexities make this apparently routine work highly non-routine. The work practices of these technicians require constant improvisation. Additionally, MRO in industry is even more complex due to the fact that industrial facilities are complex systems of buildings, infrastructures and machines [5].

2.1 Terms, Classifications, Policies

To minimize risks and to increase productivity and efficiency, there are different investigations and policies from the field of operational research. Paz et al. [6] summarize two broad classes of policies for maintaining and reliability. First: Policies for reducing the frequency of failures, e.g. (a) preventive maintenance, (b) early equipment replacement, (c) predictive maintenance and (c) overbuilding of equipment. Second: Policies for reducing the severity of failures, e.g. (a) speeding the repair task

by modular equipment design and (b) providing alternate output during repair by alternate job routings, standby machines or buffer inventories.

Following [6] it is also useful to classify maintenance on the basis of when the work must be done: (1) emergency, breakdown or reactive maintenance (RM) is work that must be done immediately; (2) routine or planned maintenance (PM) is work that must be done in the finite, foreseeable future; and (3) preventive maintenance (PM) is work that must be carried out on a planned schedule. However, the difference between PM and RM is more of a continuum than a dichotomy in many situations and there is a need for further investigations in the complex relationship between PM and RM to integrate breakdown and preventive maintenance. For this it is necessary to understand the nature of RM from a practitioner's perspective which is the aim of this contribution.

2.2 MRO in Research

There are already some interesting contributions in the literature of CSCW around MRO. Kovalainen et al. [7, 8] performed a large scale investigation on the work of operators in Finnish paper mills. They tried to understand the role of Organizational Memory for the support of communication between workers of different shifts. They introduced an electronic diary and analyzed the conversations articulated in this diary and compared the changes in practice with the usage of a prior paper based diary.

Legner and Thiesse [5] focus on Ubiquitous Computing to support the facility maintenance at Frankfurt Airport. Due to the strict fire-prevention policies the requirements for a standardized and controllable process regarding to accuracy, completeness and documentation they introduced RFID to support PM in this highly demanding field. However, they do not focus on RM; but for a successful and efficient MRO work they identified three criteria which are also relevant in the field presented and described in this paper: (1) precise and timely information on the objects to be maintained, (2) real-time transfer of information on critical incidents, and (3) fast access to the knowledge and means necessary to overcome problems caused by the incidents. There are already research investigations in designing technologies to support maintenance work but under different perspectives and different aims. Lampe et al. [9, 10] worked on the potentials of Ubiquitous Computing technologies to support the maintenance work on airplanes by introducing a RFID-based asset management system. However, they do not analyze the work of maintainers itself, they just offer a tool to deal with the concrete problem of managing and tracking tools because losing a tool in an airplane's engine is dangerous.

Though, from the management perspective there are some approaches to create policies to shift maintenance from a reactive to more preventive and predictive type of intervention to minimize the unpredictable and risky character of RM. Nevertheless, there are still unpredictable and unforeseen breakdown situations maintenance workers have to deal with and always will be. Through the increasing requirements regarding availability and reliability of the technical equipment in production facilities, further investigations are necessary especially in the support of reactive maintenance (RM). To understand this weak structured, knowledge intense and highly cooperative work, insights of this particular practice are necessary to provide appropriate technological support. From a theoretical perspective Phelps and Reddy [1] reflect on the

critical role boundary objects play in construction project teams. In this domain, boundary objects extend beyond their traditional role as information artifacts used to communicate between teams to serve a more influential role as guides for team collaboration. The theoretical concept of ‘boundary objects’ in those settings has been used in several former contributions to provide a better understanding how communication and cooperation is organized around relevant information artifacts (c.f. [1, 2, 3, 11, 12]).

2.3 Field, Methods and Theoretical Grounding

To design technologies to support reactive maintenance (RM) work in the complex field of production facilities it is necessary to understand how maintenance workers work and how they deal with the complex and time critical task of breakdown intervention. To get a deep understanding and to prepare the design and development of an appropriate technology the author of this paper started with an ethnographic phase based on a Grounded Theory approach [13, 14]. To collect the empirical data the author conducted three weeks of participant observations in a MRO department of a medium enterprise in the domain of plastic moulding for automotives, two weeks in the morning shift and one week in the late shift.

The collection of data and the documentation of the observed practices were subject to several restrictions. Because of the high pressure of competition in this domain and lack of trust in the beginning of the observation, the management of the company prohibited documentation with photos, audio- or video recordings. Therefore, the documentation was limited to field notes and the creation of a successive narrative documentation in form of a diary to capture the observations and experiences the author made. Observed dialogues were reconstructed immediately in every break after they occurred and written down as a field note [3]. Every day, after the end of the shift, the field notes have been integrated with and supplemented by protocols based on retrospections and memories of the day [13].

Upcoming questions during the observations have been made to a subject of discussions with local actors in ad-hoc interviews in coffee- and cigarette-breaks. The observations were rather passive in the beginning but after a few days of becoming acquainted with the workers in the department, the observations developed towards a more active participation in terms of assistance of maintenance workers and self-responsible conduction of simple tasks like running errands, cleaning a machine during PM activities etc. After two weeks of participating in the company, the author was allowed to make some photos, but only under permission of the head of the MRO department. The result of the documentation is a series of 44 photos and 42 pages of handwritten pages of field notes in the diary.

It was possible to get an insight of the used documents in the department. There were several individual paper-based handwritten documents, a cabinet with folders for each of the 54 machines containing the complete documentation including technical drawings for hydraulics, electrics (place is marked with (d) in Fig. 1). The folders also contain the list of spare parts for each machine with order codes. There was also a network drive containing all digital representations of hydraulic and electric drawings as unsearchable TIF files named and ordered by the IDs of the machines accessible from all PCs in the MRO department. During the observations the field notes have

been successively analyzed to identify patterns in the work routines to build up categories. The results have been combined with the findings from the document structure analysis [13]. These categories provide a mask for successively sharpen the focus of the observation on the aspect of boundary objects, location and locality of machines, persons, tools and information resources in the work practice of the actors in the MRO department.

Additionally, the analysis was supported by researchers from HCI, CSCW and social-sciences within three interpretation workshops to provide external perspectives on the collected data. The interpretation sessions were held after each week of observation. Each session was structured in three cycling phases. The observer created a report and a presentation about the observations he made in field. During presentation the scientists of each discipline wrote down open questions which were answered by the observer after the presentation. The purpose of this technique was to gain the observer's retrospection and to enrich the field notes with additional data. Questions which could not be answered by the observer were included into the next observation phase. The sessions were also used to recognize patterns in the observed work to sharpen the next phase of observation. After finishing three weeks of observation, the collected material has been analyzed by the same group to finish the category-building process and summarize the findings.

3 MRO in the Domain of Injection Moulding

The domain of automotive production is a cornerstone of the German industry and in particular the domain of plastic moulding is an important sub-domain for subcontracting for the automotive production and assembly industry. The subcontracting conditions mostly are based on just-in-time logistics. Therefore, breakdown situations have a critical impact on the reliability in those tight contracting relationships. Within that setting, technological support for the MRO activities seems to be important to reduce the risks of breakdowns and especially to react fast and effectively on unpredictable breakdown situations.

3.1 The ‘Automotive Plastics Inc.’

The participant observation took place in the MRO department of a small and medium enterprise (SME) in the domain of plastic molding. In this paper the enterprise is called “Automotive Plastics Inc.” which is not its real name. The company offers concept and series development of automotive interiors with focus on kinematic solutions like ashtrays, cup holders, sunroofs etc. The enterprise as a whole has approximately 1200 employees and the MRO department hosts 7 electricians, 1 mechanician, 6 mechanicians, 1 lacquerer and painter, 2 office workers (assistance of head and dispatching) and 1 leading engineer (head of the department). The members of the MRO department are responsible for maintaining 54 injection moulding machines, the corresponding infrastructure for cooling, heating, granulate material logistics and the facility, power- and energy-management.

As illustrated in Fig. 1. (left) the department resides within an own building on the area of the company (MRO). The production, were all 54 injection moulding

machines reside, is distributed over three different factory buildings (KV1-3). The MRO department is responsible for managing the spare part storage which is located in the basement of the KV1 next to the plastic granulate drying facilities, which are also located in the basement.

The MRO department building is divided in several areas itself: (a) marks the office of the head of the department. From this office the leading engineer organizes the division of labor, creates schedules of planned overhauls and calibrations in tight coordination with the head of production, leads meetings of the department, acts as an intermediate between the head of production and the MRO department. During the time of observation the head was involved in planning of investments in new facilities. Therefore he was not available at least half of the day. The neighbor-room (b) is the office of the two dispatchers. The dispatchers are responsible for receiving incoming repair claims coming from the production department. They are also responsible for ordering spare parts. Both rooms are directly connected by a door which stands open most of the time except in case of meetings in the head's office. Therefore, if the head is working in his office, he is very aware of repair claims reaching the dispatchers via phone. The group of mechanicians usually works in the area marked with (d). They have a massive workbench and the walls around it are completely covered with tools they use to work with. Nearby (e) the electricians and the mechatronician have their own workshop separated with glass-walls from the rest of the MRO department. Within this room there are many cabinets for special spare parts and components for the work of the electricians. There is a workbench in the middle of the room where every electricians and the mechatronician has its own desktop-like area. In the area of (f) there are cabinets and shelves for screws and several small parts which are needed by all MRO members.

3.2 Existing Structures, Routines and IT Support

To organize the different maintenance, repair and overhaul tasks in terms of internal cost allocation the observed MRO department uses a SAP-PM module (SAP-Planned-Maintenance) to manage each maintenance and repair task and process. In summary, there are 6 desktop-PCs in the maintenance workshop with TFT display, mouse, keyboard and unrestricted Internet access. They are running a SAP-PM Client, a browser (Internet-Explorer) and unrestricted Internet access.

Regarding the classification in section 2.1 the observed MRO work is divided in two different categories: Planned (or Preventive) maintenance (PM) and Reactive Maintenance (RM). An example for PM-activities is the routinely performed calibration of the injection moulding machines which is important for the accuracy and quality of the products. These tasks are fully integrated within the production plan and structured by a step-by-step instruction which is part of each machine's documentation. The head of the production coordinates these regular tasks together in cooperation with the head of the MRO. He constitutes teams of electricians and mechanicians (2-3 people) and gives them the order when and where the calibration has to take place.

The official process of dealing with an unplanned breakdown is organized as follows: The operator of a machine recognizes failures or quality loss of the output of an injection moulding machine or she is not able to operate the machine anymore because the machine stops and switches into an error state. Then she asks her fitter to

help her solving the problem. If they are not able to fix the problem autonomously, the fitter calls the MRO department. The dispatcher registers the breakdown and clarifies how urgent the machine is needed within the production plan. She creates a new repair task within the SAP-PM by typing a short description of the case, assigning a responsible MRO worker, estimating the effort in hours and printing out a documentation- and time-sheet (c.f. DIN-A4 sheets in Fig. 2.3). After that, the dispatcher commits this sheet to the responsible maintenance worker. In urgent cases the maintenance worker has to interrupt her current task to directly start with fixing the actual breakdown. After she successfully finished her work, she has to do the documentation of the finished case. She has to use the time sheet to fill in the time needed, a short description of what she did, which spare parts she took from the store and finally has to sign for confirming the accuracy of the information. After that she hands back the complete time sheet to the dispatcher. The dispatcher enters the information from the sheet into the SAP-PM, orders spares removed from the store and finishes the task. The collected data is used later on to calculate the internal cost allocation and calculating and anticipating the needed man-power of the MRO department.

4 The Observed Practice

In the following we provide some observed examples of the complex work of diagnosing, intervention and documentation which provides a good basis for identifying weaknesses and gaps in the current processes and gives an insight of the observed MRO work. Especially *cooperative aspects* around *relevant objects* and their *locality* and *location* during the *diagnosis phase* are taken into consideration. Consequently the documented cases focus on the “first-responding” like practice of RM within the MRO department - which is often embedded into PM activities.

4.1 Case 1: “Just” a Broken Ventilator Grid

A new repair-claim arrives at the MRO department. There is a broken ventilator grid at a machine from the vendor “Arburg” in KV1. Due to some safety restriction it is not allowed to run the machine as long as it is not fixed. Therefore, the responsible fitter (Mark) from the production department calls the department and articulates a repair-claim. The dispatcher (Paul) creates a repair task in the SAP-PM and assigns it to an electrician (Peter). Paul states that there are some grids left in the spare store. Peter walks to the machine and measures the broken grid to identify the right form factor and makes some notes in his personal diary. After that he walks into the spare store to get the right grid. After 10 minutes of searching Peter states that there are lots of grids but not one that fits. He then walks back to the machine and describes the problem to the fitter Mark. After that he calls the dispatcher Paul with his cordless telephone (DECT). The dispatcher orders the missing spare part and states that is will arrive by no longer than the day after tomorrow, maybe earlier. Mark blocks the machine in the production plan and alternates to another machine.

In the meantime Peter receives a call from a colleague (Josh) who is working at another ‘Arburg’ machine in the ‘KV2’ which is 15 years old. Arriving at the machine the problem has reached another MRO worker - a mechanician (Andy). He arrives at

the same time as Peter. Josh immediately starts to explain the problem with the handling. After this short introduction all three MRO workers in place try to reconstruct the complete repair history to figure out if there were similar cases in the past and who was responsible. During the discussion everybody points to several components of the machine, walk around it, watching the handling from different perspectives. After approximately 15 minutes of diagnosing and discussion Peter walks back to the MRO department to fetch the documentation of the machine. 10 Minutes later Peter is back from the MRO department. He had some problems finding the right folder with the documentation.

Then, another electrician colleague named Kevin calls Josh. He has received a breakdown of the central drying facilities. Josh states that he has no time at the moment to help with that. Josh asks Peter to walk down in the basement to help Kevin. After that Peter and Andy leave Josh to visit Kevin. In the basement - nearby the spare parts store - Kevin waits by the main drying unit of the KV1. The drying unit is responsible to prepare the plastic granulate to transfer it from the delivery state into production state. The material has to be very dry (dew point under -60° Celsius) to ensure high quality standards. He states that the head of production reported problems with some products (bubbles and foam) and the production data acquisition confirmed

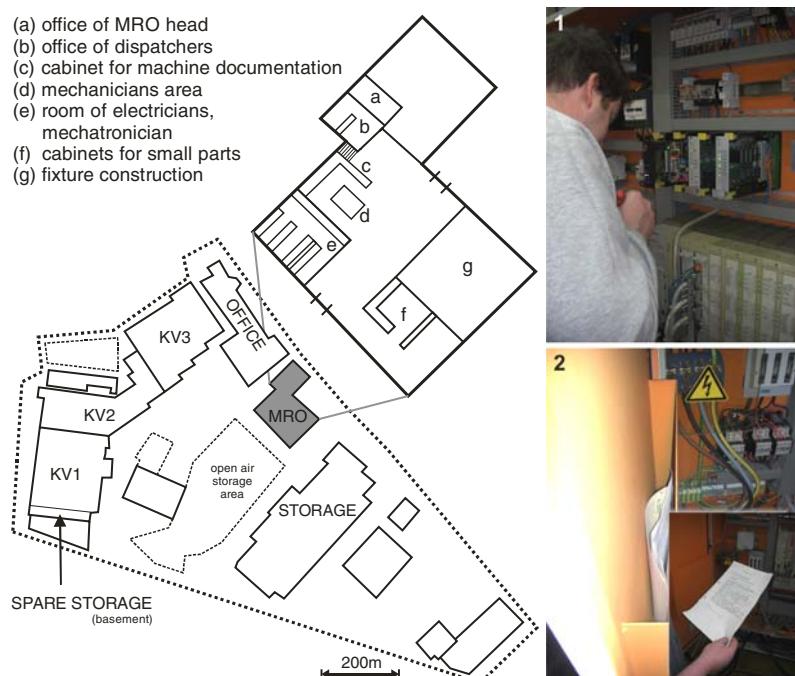


Fig. 1. left: Area of the Automotive Plastics Inc. and detailed view on MRO department building | right: . (1) Josh, cleaning cartridges (2) step-by-step instruction for a complete reset of the PLC found in a compartment for documents in the switchgear cabinet's door.

the suspicion of humid material. Humid material in a centralized material management could cause a complete stop of production in KV1 which has a deep impact on the just-in-time delivery to the customers. After a short discussion everybody recommends exchanging the heating elements of the drying unit. After fetching some new heating-elements from the store nearby Peter leaves the basement and goes back to the MRO department to have a break. After 10 minutes he calls Kevin, who is still in the basement. He figures out that Kevin seems to be still very nervous because after the heater element exchange the temperature of the drying unit does not increase but does not decrease as well - which is probably a good sign.

After the call Peter makes some notes in his diary. He explains, that they have to document their work after completion and hand over the filled in time sheet to the dispatcher. But after the break he has to start working on a recalibration of a machine in KV2 and therefore he has no time to do the documentation work directly after the tasks. He states - while pointing on a stack of timesheets on his workbench - that he collects the time sheets until he has to work in the late shift. From his point of view, late shifts are not that busy. To recall the relevant information he takes some notes in his personal diary (cf. Fig. 2.3). He states that all his colleagues have the same strategy and some of them do not even make notes. Finally, Josh returns the MRO department. He looks satisfied and reports correctly running handling at the “Arburg”. There was corrosion at a plug which hindered vertical movements of the handling.

4.2 Case 2: Diagnosing a Machine-Breakdown

A fitter from KV1 calls Josh directly at his phone without calling the dispatcher. Josh states that he knows him very well and he calls him regularly. He reports a breakdown of a machine from the vendor “Demag”. The fitter tries to explain the problem by describing the current state of the machine and the error-code prompted at the display. Josh quits the call by promising that he will come over to take a look at that. On the way to KV1 he meets the head of the department, Mr. Sheppard. He describes him the actual breakdown situation and the name of the machine. Mr. Sheppard has already heard from the new breakdown because he comes from KV1. He states the case is trivial and easy to fix. He worked on the 12 year old machine several times and remembers the error pattern very well. The solution is to remove all cartridges of the PLC control unit, clean them and treat the contacts with a special spray. After that he should reboot the machine. It would work again “*...like if it was new.*”

Then, Josh meets the mechanician Andy and tells him about the new breakdown. Because Andy likes the older “Demag” machines in the park and Josh is much younger and works just half a year with the company he accompanies him on his way to the machine. The fitter is already waiting and begins to explain the current issue coupled with pointing at the display and recalling the situation in which the machine stopped working. Josh opens the switchgear cabinet and directly puts his head into it. He states that this is a typical way to identify fused electrical components. He smells nothing unusual and opens the doors of the cabinet completely.

Meanwhile, Andy started to move the mould manually by switching the machine into the manual mode. He is irritated because nothing happens. He takes a look at the display and sees some error codes (c.f. Fig. 2.1/2/4.). The responsible fitter states that the error codes do not say that much about the state of the machine. The display

shows error codes all the time but works without any problems. Kevin passes by and asks if he could help. Josh shuts down the machine and starts to apply the workaround of Mr. Sheppard. After removing all cartridges (c.f. Fig. 2.1) he figures out that one of them is darker than the others and smells a little bit smoky.

After that Kevin, the fitter and Andy started a discussion about the breakdown history of the machine and the proposed workaround of Mr. Sheppard. Andy is not sure if Mr. Sheppard has enough expertise and experience to make helpful suggestions. Josh joins the discussion and shows the probably fused cartridge. Kevin smells at it and is not sure if it is fused. Josh cleans it, uses the spray and puts all cartridges back into the PLC control. He calls the dispatcher and asks him if there is an appropriate replacement part in the store. The dispatcher is not sure, because this machine is relatively old compared with the rest of the park. Andy walks downstairs and tries to find a new cartridge.

The fitter (who is part of the production- and not of the MRO-department) states that approximately one year ago a maintenance worker did a complete reset of the PLC unit. As far as he remembers, it was the head of the MRO, Mr. Sheppard. Kevin calls him and asked him about the reset. Mr. Sheppard remembers the reset and states that there is a paper with the instruction. But he is not sure where it is. He proposes to have a look into the documentation folder of the machine which resides in the related cabinet in the MRO department.

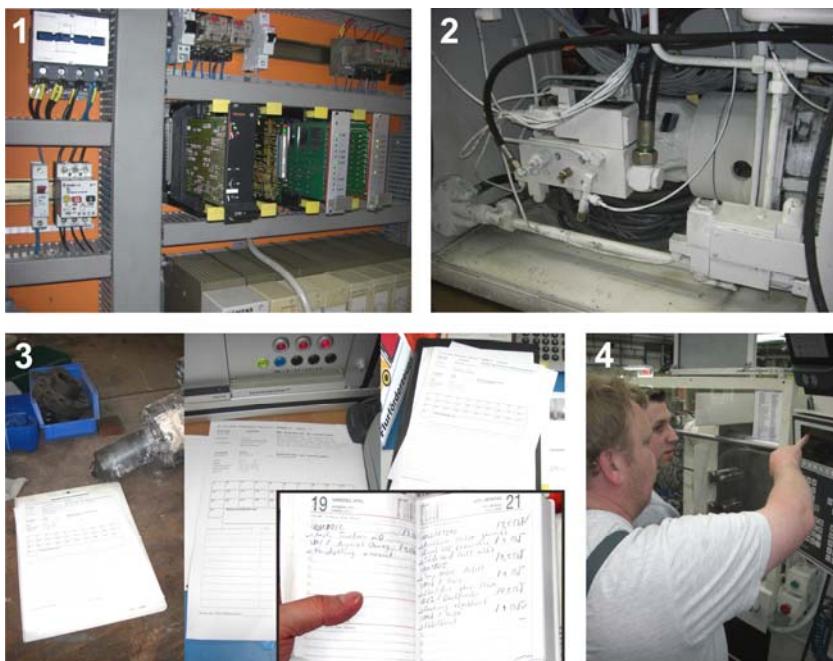


Fig. 2. (1/2/4) Diagnosis as a cooperative and discursive practice in place, with directly referencing to relevant objects between different experts (mechanician in front and electrician behind) | (3) A personal diary containing a list of tasks with the needed time in hours. Stacks of uncompleted time sheets in the MRO department.

Kevin walks back from KV1 to the MRO department and starts to look for the right folder which takes some time. After he finds it he goes back to his colleagues and they start to search the folder for the sheet of paper. After 10 minutes of browsing the folder they surrender and Kevin calls Mr. Sheppard again. He promises to come around and as he joined the others he pointed to a compartment on the inner surface of the switchgear cabinet's door (cf. Fig. 1.2). Mr. Sheppard tells everybody how this paper was created several years ago. The reset to factory defaults was not part of the original documentation of the machine. In those days, the instructions were dictated by an external service man via phone and written down with a typewriter. Mr. Sheppard is sure that the copy they found in the cabinet is the last copy the company has and he recommends making a new copy and putting it into the documentation folder. Josh starts immediately resetting the machine by following the step-by-step instructions. After that, the machine still displays error messages and does not work. The fitter points at the electro-static hydraulic-oil-filtering-device which is plugged into the machine. This is a PM task the fitters have to perform autonomously. He states that older PLCs sometimes crash if a filtering device is plugged into a machine. After the fitter has switched off the filtering device, Josh reboots the machine again. There is no error message anymore and the fitter installs the actual program again. The machine works normally, in automatic and manual mode. Josh and Andy go back to the MRO department and discuss the electrostatic effect on older PLCs. They state that this was definitely something new and a lesson learned.

5 Findings

The descriptions of the two typical cases illustrate the complexity of the work in a MRO department in the addressed domain. In the following the model of a standard RM process is used as a basis to structure the observed MRO practice. Within each phase relevant groups of actors are identified. The analysis focuses on broken machines on-site and their role as ‘Boundary Objects’ in the cooperative process of repair and overhaul (c.f. [1, 2, 3]). From the findings implications for a design are introduced to provide a guideline to design information technology to support the cooperative MRO work around those Boundary Objects – involving the actual groups of interest and additional new groups in future.

5.1 Breakdown and Repair Claim

Especially small and medium enterprises are social systems where people know each other very well. Hence, in case of a machine’s breakdown, there are different ways of getting help: the first and official way is calling the dispatcher. This strategy has a “fire and forget” character and the dispatcher is responsible to find the right maintenance worker. A dominant observed practice is directly calling somebody you already know. In this case the dispatcher does not know something about the breakdown, therefore, the time sheet has to be created and filled out afterwards. A similar way to get help is to ask a maintenance worker who walks by or works at another machine nearby. Repair claim and diagnosis is often an interwoven process. The diagnosis often starts during the phone call before the maintenance worker has reached the

defective machine because the responsible operator or fitter in-place starts to talk about the new case to isolate possible error sources. MRO workers avoid those situations by quitting the calls. As observed in both cases MRO workers prefer to visit the machine and talk face to face and have look at the affected objects. In the early phase of breakdown and repair claim maintenance workers which are not directly involved often feel responsible for the actual case without an official assignment of the dispatcher. If a maintenance worker needs help and has to act in time critical conditions, he seeks for the help of other maintainers without requesting it from the dispatcher. Therefore lots of repair claims also come directly from maintenance colleagues. Maintainers seek the help of those colleagues which seem to have the appropriate profession and/or experience to make a helpful contribution. They try to find out, which colleague has worked with the affected machine in the past.

In summary, actors from various involved parties use the broken machine as a meeting-point for the discussion of possible reasons for and consequences of the current breakdown. This forms the stage for the next phase, the diagnosis.

5.2 Diagnosis

The main diagnosis phase usually starts when the maintenance worker reaches the defective machine. Local actors like operators and fitters are important initial counterparts in early stages of the diagnosing process. Regarding the involvement of the different actors in place the isolation of the error source is a highly cooperative, creative and improvisational practice. In all observed breakdown situations maintainers asked their direct colleagues and the responsible operators and fitters to have a look at the problem and to take part in a discursive diagnosing conversation. By observing and listening to the discussions it is noticeable that every observed discursive diagnosis situation is based on the reconstruction of the past. The history of a machine seems to be very important for the isolation of the source of error. Maintenance workers experience that the older a machine is, the probability increases that the actual breakdown happens before. Hence, discussions often are led by the challenge to reconstruct the ‘machines life’. The discursive reconstruction is strongly connected to the physical presence of the broken machine. This was observable in the integration of the machine’s parts into the discussion by pointing and touching.

To increase the accuracy of the reconstruction it is very important to include other maintenance workers into the discussion - especially the older and more experienced. Relevant information is often hidden in aspects of what happened in the past around a particular machine and who was involved. It was never observed that a maintainer browsed the SAP-PM for reconstruction purposes which is caused by the weak documentation (as described later on). Especially older machines have a history which often blends over into a world of myths, e.g. stories about persons who are no more part of the team since years. In comparison to the effort of fixing a particular problem diagnosis work is very time intensive. Diagnosis work can tie up 3 persons for 30 minutes just for isolating the possible error source. Dependent on the profession, maintainers use all senses to isolate the possible defective component of a machine. Mechanicians mostly switch the machine into manual mode and move the mould back and forth. While doing this they listen to the sounds of pumps, gears and joints. They touch several components to feel the vibrations. Everything which seems to be

uncommon is a possible source of error. Electricians often use their nose to smell if electrical components are fused (cf. [15]).

5.3 Intervention

It was observed that repairing is often an integral part of the diagnosis work. Looking at Case 2, after isolating possible error sources the maintenance workers used spare parts and exchanged possible defective components to exclude possible error sources systematically. If the defective part is found the new part resides into the machine and the intervention is finished successfully. Therefore, spare parts play an important role in the intervention phase. The selection of the right spare part is determined by the spare part list which is part of the machine's documentation folder. This paper based folder is an important abstraction of the concrete machine and often provides alternative views on it. Due to their complexity and level of abstraction the construction plans of the machines are exclusively used by the MRO workers and are living documents; i.e. they are annotated by MRO workers during PM and RM activities. They are also used as boundary objects to support the communication and the collaboration between different professions within the MRO department. However, the documentation folders are experienced as an inconvenient medium, hard to find and hard to browse. The use of the documentation folder sometimes is concurrent: the documentation folders in the MRO department are needed by different MRO members. The dispatcher needs the replacement spares list to order new parts if they were taken from the store. The maintenance workers need the documentation for PM activities like recalibrations and for diagnosing and intervention in RM. In this case sharing this paper based resource leads in a centralized storage in the MRO department which causes time intensive and exhaustive 'legwork' (c.f. section 3.4 Documentation).

5.4 Documentation

The documentation within the SAP-PM generated time sheets often is quite poor. Studying several hundred cases within SAP-PM reveals that more than 80% of the documentation of the performed repair activities consists of only one sentence: "*Maintained and repaired.*" As observed in the first case the maintenance workers have a batching strategy of filling out the time sheets. One week between the performed task and its documentation is not unusual. From an analytical point of view there are several interdependent reasons for that phenomenon: The SAP-PM has a process-oriented structure. The time- and documentation-sheet has to be filled out at the end of a successful performed repair claim. As illustrated in the first case sometimes spare parts are not available from the store. In consequence, the MRO worker has to postpone finishing the task until the new part arrives. Only when he has got the parts and finished the task he is able to fill out the form on the sheet (date and time of breakdown, time needed, date and time of finishing the task, used spare parts, affected machine, etc.). From a process perspective several repair tasks run in parallel for days which leads to a batch-strategy documented in the first case. Another reason is a motivational problem: maintainers have no direct benefit from providing a detailed documentation. The SAP-PM database is not comfortably searchable and the access to the database is only possible by leaving the place of the defective machine or facility and walking back to the

MRO department. In consequence distances of walking from KV1 to the MRO department and back can summarize up to 500m by walking.

6 Implications for Design

From the different perspectives of the involved actors machines or any other equipment in the observed setting have two facets. First, from the perspective of the production department, machines are necessary resources for the production processes and need to be always up and running for reliable production planning in tight just-in-time logistics. That is, they act as an infrastructure for productive processes. On the other side, from a MRO perspective, machines, their (spare) parts and their abstract representations (e.g. the construction plans) are the centerpiece of consideration. They are not infrastructures for a particular purpose; they are the artifact to work on. In that role they act as boundary objects by enhancing the coordination and communication between different stakeholders in breakdown situations.

Star et al. [16] define ‘Boundary Objects’ as “... objects which are both plastic enough to adapt to local needs and constraints of the several parties employing them, yet robust enough to maintain a common identity across sites. They are weak structured in common use, and become strongly structured in individual-site use. They may be abstract or concrete. They have different meanings in different social worlds but their structure is common enough to more than one world to make them recognizable means of translation. The creation and management of boundary objects is the key in developing and maintaining coherence across intersecting social worlds.”

Especially the diagnosis and isolation of defective parts is a highly cooperative and discursive task where the broken machine and its parts in certain places serve as physical meeting points for different involved groups of stakeholders. They are used as mental ‘drawing tables’ for information exchange between different groups in the organization and outside. All involved groups gather around them to exchange information about their experiences to make contributions to solve the current problem. Following this, machines and spare parts serve as boundary objects to enhance the development of a common understanding between all involved actors around a machine breakdown. To enrich physical objects in the observed domain to emphasize their role as boundary objects in MRO work the following guidelines are derived from the findings.

Support documentation work

Documentation is an important basis to build up the breakdown history of each machine, facility or infrastructure the MRO department probably has to deal with in future. In the current state the documentation work is experienced by the MRO workers as an extra workload, performed out of the concrete working context and - from the workers perspective - provides no benefit for the practice. However, as illustrated in case 2, maintenance workers already use paper based handwritten notes to build up personal maintenance histories (c.f. [17]). They are weak structured informal notes but never the less accessible and useful on-site. Referring to the introduced theoretical framework of boundary objects, the documentation history should be accessible for every involved actor in-situ. The actual documentation-work should be performed

during all phases of intervention and the complete documentation history should be directly associated with the relevant objects. IT support should extend the relevant physical objects to provide a surface, plastic enough to enable every involved actor to express whatever they think is useful for others in the future. The challenge probably still is motivating the workers to document their work - especially if there is no direct benefit for them. As observed in the field, most of the MRO workers are motivated to document their work because they take advantage from those personal notes in future. 'Publishing' those notes and associating them to the objects could be an important first step. The motivational challenge could also be faced by partially automating the documentation of involved persons, used spare parts, date and time of the breakdown and the completion of the task. Ubiquitous computing technologies like autoID systems could be used to identify involved actors and used spare parts around machines.

Bridging between the machines and documentation

To support maintenance workers finding relevant information repositories have to be searchable and browseable from the place where the action is: the defective machine or facility. As described in the both cases - especially during the diagnosis phase - the responsible MRO workers walk around the machines, point to relevant objects and try to reconstruct the breakdown history of the machine. In the current state of IT support the relevant information about the history of the machine is stored case by case and process-oriented in the SAP-PM system. MRO work is an object-oriented work; strongly rooted in the physical, therefore a layer of 'bookmarks' providing a direct access to relevant information attached at the affected objects would be useful to improve the often speculative and error-prone diagnosing work of reconstructing the past. Those 'markers' or 'hooks' should be brought out by the MRO workers themselves while they are coping with a breakdown situation because they have the experience and the expertise to associate the right object with the relevant information.

Provide access to local and remote experts

The third and probably most important aspect is finding the right local experts who have the knowledge and experience to help in a particular breakdown situation (cf. [18, 19]). The identification and documentation of involved persons in certain breakdowns is an important aspect of supporting indirectly the MRO worker's work. Support for finding the right colleague for a discussion about the current problem in the diagnosis phase is an important contribution for a more efficient and effective diagnosis. Because the diagnosing phase is time consuming efficiency and effectiveness in this phase is a promising contribution to reduce downtime of production-critical machines. To involve other potential stakeholders this locally created and stored information should be accessible from other places to enable MRO workers to browse the whole information space for similar or related cases in other places. In very complicated and persistent breakdowns it is also possible to enhance the cooperation with external experts, e.g. from a particular vendor, by giving them access to a machine's history. In this case, the information space around augmented machines could also work as a boundary object to involve external experts and provide all relevant information about the history to them, too (c.f. Fig. 3 'external actors').

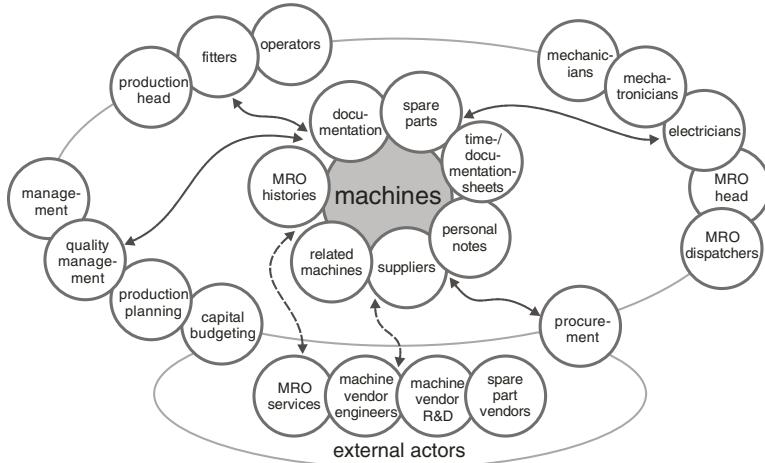


Fig. 3. Machines as boundary objects for supporting collaboration between related stakeholders in the organization of MRO and MRO related activities

7 Conclusion and Future Work

The current IT support in the observed setting is too office-, process- and management-oriented to fulfill the information needs for the operative MRO workers. The observations reveal that RM activities are still an important part in MRO. In current research - especially against the background of the increasing application of embedded computing technologies - there are only few contributions which give practical insights of this work and discussing them against paradigms of pervasive and ubiquitous computing. The enrichment of physical objects with embedded ubiquitous information technology has two key characteristics which are highly relevant in the addressed domain: First, relevant information moves in places where the need of information occurs - embedded into the existing practices. Second, the same information, created and explicated in-situ gets available independent from the location.

The applied qualitative ethnographic methods revealed some important findings which lead to implications for design of IT support in the addressed domain. Regarding their special characteristics we propose the introduction of Ubiquitous Computing technologies to embed relevant information into the locally contextualized practice around objects on-site especially to support the cooperative and discursive reconstruction of the breakdown history in-situ. Augmenting the physical objects in place could provide a plastic semi-virtual surface where maintainers and other involved groups exchange information. This would bring the defective machines into the role of boundary objects that mediate between all involved groups of actors in breakdown situations. Due to the limited observation time in-place and the limitations regarding the gathering of empirical data the next step is to build first design sketches and introduce them in the observed practice with the aim of gathering more data about this indeed very complex domain and to observe early signs of appropriation to inform the further design. E.g. autoID-technologies could be a first appropriate means to enable

the MRO workers to annotate affected objects with links to information like involved persons, relevant spare parts and folders in the machine documentation cabinet (cf. [2, 3, 4]). The next evolution of this approach will be the introduction of ad-hoc deployable and interconnected sensor nodes to automatically gather additional information about a machine's history like deviating vibrations and temperatures to both enhance the prediction of breakdowns and to support the reconstruction of the past. Further investigations have to consider if autoID systems and wireless sensor networks work in environments with lots of metal or if they sustain on temporally hot surfaces.

References

1. Phelps, A.F., Reddy, M.: The influence of boundary objects on group collaboration in construction project teams. In: Proceedings of the ACM 2009 international conference on Supporting group work - GROUP 2009, pp. 125–128. ACM Press, New York (2009)
2. Star, S.L.: The structure of ill-structured solutions: boundary objects and heterogeneous distributed problem solving. In: Distributed Artificial Intelligence, vol. 2, pp. 37–54. Morgan Kaufmann Publishers Inc., San Francisco (1989)
3. Lutters, W.G., Ackerman, M.S.: Beyond Boundary Objects: Collaborative Reuse in Aircraft Technical Support. Computer Supported Cooperative Work (CSCW) 16(3), 341–372 (2006)
4. Orr, J.E.: Talking About Machines: An Ethnography of a Modern Job (Collection on Technology and Work). Cornell University Press, New York (1996)
5. Legner, C., Thiesse, F.: RFID-Based Facility Maintenance at Frankfurt Airport. IEEE Pervasive Computing 5(1), 34–39 (2006)
6. Paz, N.M., Leigh, W.: Maintenance Scheduling: Issues, Results and Research Needs. International Journal of Operations & Production Management 14(8), 47–69 (1994)
7. Auramäki, E., Robinson, M., Aaltonen, A., Kovalainen, M., Liinamaa, A., Tuuna-Väiskä, T.: Paperwork at 78kph. In: Proceedings of the 1996 ACM conference on Computer supported cooperative work - CSCW 1996, pp. 370–379. ACM Press, New York (1996)
8. Kovalainen, M., Robinson, M., Auramäki, E.: Diaries at work. In: Proceedings of the 1998 ACM conference on Computer supported cooperative work - CSCW 1998, pp. 49–58. ACM Press, New York (1998)
9. Lampe, M., Strassner, M.: The Potential of RFID for Moveable Asset Management. In: Proceedings of the 5th International Conference on Ubiquitous Computing, Workshop on Ubiquitous Commerce. Springer, Heidelberg (2003)
10. Lampe, M., Strassner, M., Fleisch, E.: A Ubiquitous Computing environment for aircraft maintenance. In: Proceedings of the 2004 ACM symposium on Applied computing - SAC 2004, pp. 1586–1592. ACM Press, New York (2004)
11. Carlile, P.R.: A Pragmatic View of Knowledge and Boundaries. Organization Science 13(4), 442–455 (2002)
12. Martens, A., Hambach, S., Lucke, U.: Multi-perspective Cooperation Based on Boundary Objects. In: 2009 Ninth IEEE International Conference on Advanced Learning Technologies, pp. 476–478. IEEE, Washington (2009)
13. Glaser, B.G., Strauss, A.: The Discovery of Grounded Theory: Strategies for Qualitative Research. In: Aldine Transaction, Piscataway, NJ, USA (1999)
14. Strauss, A.: Work and the Division of Labor. The Sociological Quarterly 26(1), 1–19 (1985)

15. Suchman, L.: Embodied Practices of Engineering Work. *Mind, Culture, and Activity* 7(1), 4–18 (2000)
16. Star, S.L., Griesemer, J.R.: Institutional Ecology, ‘Translations’ and Boundary Objects: Amateurs and Professionals in Berkeley’s Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science* 19(3), 387–420 (1989)
17. Luff, P., Heath, C., Greatbatch, D.: Tasks-in-interaction: paper and screen based documentation in collaborative activity. In: Proceedings of the 1992 ACM conference on Computer-supported cooperative work, pp. 163–170. ACM, New York (1992)
18. Pfeffer, J., Hinds, P.: Why Organizations Don’t “Know What They Know”: Cognitive and Motivational Factors Affecting the Transfer of Expertise. In: Ackerman, M., Pipek, V., Wulf, V. (eds.) *Sharing Expertise: Beyond Knowledge Management*, pp. 3–32. MIT Press, Cambridge (2002)
19. Reichling, T., Veith, M.: Expertise sharing in a heterogeneous organizational environment. In: Gellersen, H., Schmidt, K., Beaudouin-Lafon, M., Mackay, W. (eds.) *Proceedings of the ninth conference on European Conference on Computer Supported Cooperative Work*, pp. 325–345. Springer, New York (2005)

Automatic Assessment of Cognitive Impairment through Electronic Observation of Object Usage

Mark R. Hodges¹, Ned L. Kirsch², Mark W. Newman³,
and Martha E. Pollack^{1,3}

¹ Computer Science and Engineering

² University of Michigan Medical School

³ School of Information

University of Michigan, Ann Arbor, MI, USA

{hodgesm, nlkirsch, mwnewman, pollackm}@umich.edu

Abstract. Indications of cognitive impairments such as dementia and traumatic brain injury (TBI) are often subtle and may be frequently missed by primary care physicians. We describe an experiment where we unobtrusively collected sensor data as individuals with TBI performed a routine daily task (making coffee). We computed a series of four features of the sensor data that were increasingly representative of the task, and that we hypothesized might correlate with severity of cognitive impairment. Our main result is a significant correlation between the most representational feature and an apparent indicator of general neuropsychological integrity, namely, the first principal component of a standard suite of neuropsychological assessments. We also found suggestive but preliminary evidence of correlations between the computed features and a number of the individual tests in the assessment suite; this evidence can be used as the basis of larger-scale studies to validate significance.

1 Introduction

Cognitive impairments such as dementia and traumatic brain injury (TBI) can be difficult to detect and assess—one study showed that up to 75% of cases of dementia or probable dementia go undiagnosed by primary care physicians [1]. Additionally, cognitive ability may vary from day to day and, since therapists cannot observe patients on a daily basis, they are often forced to rely on questioning patients about their activities and to “play detective” using the answers given [2]. This paper describes work to automatically assess cognitive impairments caused by traumatic brain injury by using wireless sensors to observe individuals performing an everyday task, and then extracting features that are correlated to important neuropsychological assessments. The use of simple wireless sensors can potentially enable unobtrusive assessment on a daily basis in a naturalistic setting.

We asked individuals with TBI to make a pot of coffee and electronically observed them by placing RFID tags on relevant objects and having the subjects wear a bracelet with an RFID reader. We chose coffee making as a common

functional task that serves as a proxy for everyday activity performance. We analyzed the collected sensor data and compared it to the subjects' scores on a standard suite of neuropsychological assessments. Our analysis considered a series of four features computed from the data, which were increasingly representative of the task. We found a significant correlation between the most representational feature—edit distance from a correct task plan—and the first principal component of the assessment suite, which appears to serve as an indicator of general neuropsychological integrity. Because of the large number of tests in the neuropsychological suite, we were unable to collect sufficient data to demonstrate statistical significance between our computed features and individual tests, but we did find suggestive yet preliminary evidence of correlations, which can be used to structure larger scale investigations.

2 Motivation

The study we report on in this paper involved subjects who were being treated for traumatic brain injury (TBI). TBI is not uncommon: approximately 0.46% of Americans are hospitalized for brain injury each year and individuals aged 15-24 are far more likely than any other age group with over 0.9% hospitalized each year for brain injury [3]. Regrettably, TBI is frequently seen in wounded veterans returning from the Iraq War. Improved body armor has helped soldiers survive explosions that they might not have survived before, but many soldiers are suffering brain damage as a result of the blasts. The increase in Traumatic Brain Injury has been so dramatic that it has been called the “signature wound” of the Iraq War [4]—in one study of servicemembers arriving at Walter Reed Army Medical Center with injuries caused by explosions, 59% of the soldiers were found to have TBI and 56% of those were considered moderate or severe [5]. The existence and severity of TBI can be difficult to assess, in part because it cannot always be detected with imaging tests [6].

While our study was restricted to subjects with TBI, we anticipate that the approach we are using can generalize to other causes of cognitive impairment, notably including dementia. This is important, because the world's population is aging. In 2000, 12.4% of the U.S. population was aged 65 and older, and it is predicted to increase to 19.6% by 2030 and 20.6% by 2050. The oldest subgroup, that of individuals aged 80 and older, is expected to rise even more dramatically, more than doubling from 3.3% of the population in 2000 to 5.4% in 2030 and 8.0% in 2050 [7]. Trends worldwide are similar [8]. Dementia becomes much more common with age, affecting fewer than 1% of individuals in North America aged 60-64, but nearly 13% aged 80-84 and more than 30% of those over age 85 [9]. Without scientific advances to lower the incidence rates or the progression of Alzheimer's—the most common form of dementia—it is estimated that between 7.98 and 12.95 million people in the United States will have Alzheimer's Disease in 2050, four times the number that did in 2002 [10]. While our work overall is motivated by the challenges in assessing a wide range of cognitive impairments,

it is important to keep in mind that we have so far only conducted tests using patients with traumatic brain injury.

By using wireless sensors placed in a home environment, passive and ongoing observation and re-evaluation may be possible without disruption of the individual's life or schedule. This would allow the observation of subjects over an extended period of time to provide information about both short-term and long-term changes in impairment. Short-term changes, for example improvement caused by successful medication or treatments, or sudden degradation resulting from a side-effect of medication, could be quickly detected and acted upon. Simplifying long-term observation means subtle changes could be more easily detected and that day-to-day variation could be distinguished from long-term changes.

3 Background

3.1 Automated Detection of Cognitive Impairments

Researchers at the Oregon Health & Science University are developing several techniques for the automatic detection of cognitive impairments, including automatically observing users play a modified version of the game FreeCell. One study focuses on mouse movement during the game [11] while others focus on the subjects' performance over time, comparing it to the performance of an automated solver. Using the results, it was possible to differentiate the three mildly cognitively impaired subjects from the six others [12]. Work with several other computer games, specially created to perform assessments of cognitive impairments is underway with some promising early results [13]. They have also studied automatically monitoring mobility because slowed mobility may be a predictor of future cognitive decline. The time to answer a phone call was used to measure mobility [14], as were passive infrared detectors and several models to infer the mobility of subjects more directly as they move about a residence [15].

Research by Glascock and Kutzik used various sensors to observe activities of daily living (ADLs) in a subject's home. The output from the sensors, however, was analyzed manually [16]. Other research by Barger, Brown, and Alwan observed subjects using motion detectors to detect behavioral patterns. Although basic analysis of behavioral patterns was performed, this analysis was only loosely tied to performed activities [17]. Finally, Hoey, et al use an estimate of the subject's functionality in their system that assists a user with dementia during handwashing. This estimate is updated over time as the user completes the handwashing task. While detection of cognitive impairments was not the focus of that research and accuracy results are not given for their system, this approach could potentially be expanded to those goals [18].

Similar ideas have been used to address other conditions such as automatically observing autism using accelerometers placed on the wrists and chest [19]. Preliminary studies have also examined using toys with sensors to support assessment of a child's development [20].

3.2 Activity Recognition

Activity recognition is an active field of research that uses various sensors to monitor individuals, applying interpretation algorithms to recognize the activities they are performing. While we do not perform activity recognition in the current study—we instead assess the performance of known activities—there are clearly connections between the two tasks. Different applications in activity recognition focus on a wide range of activities. Recognizing whether a subject is moving in ways such as jumping or walking [21], identifying a user’s common destinations in a city [22], and differentiating whether a user is taking medication, making cereal, or eating cereal [23] are all examples of tasks distinguished by activity recognition systems.

Several types of sensors can be used to observe interactions with objects, such as RFID readers, motion detectors and accelerometers designed to detect object-use interactions, as well as electric current and water flow detectors [24]. In each case, the sensors measure approximations of object usage: with RFID readers, for example, proximity of a hand and an object is used as a proxy for object interaction; with accelerometers, movement of the object serves as a proxy. Several techniques have been used to analyze this data, including probabilistic methods and decision trees [24][25].

There are also many approaches to activity recognition that are not based on the analysis of interactions with objects, including the use of GPS [22], small wearable sensing platforms that have several sensors including accelerometers [26], and data-rich sensors such as video cameras or microphones [21].

4 Methodology

Our study involved subjects performing an everyday activity that could be monitored using wireless sensors. We hypothesized that patterns of errors made in the performance of such activities are associated with the severity and type of a patient’s cognitive impairment and further, that we could use wireless sensors to accurately detect those errors. We were concerned both with predicting overall neuropsychological integrity, and with identifying more specific neuropsychological profiles, such as isolated difficulties with memory, attention, or executive reasoning.

4.1 Selection of a Task

We chose to observe subjects preparing a pot of coffee using a drip coffee maker common in North America. This task was selected because it is performed by many people on a regular basis, so individuals could be assessed as they perform their daily routine. Variation in the ways that individuals make coffee is somewhat limited so patterns can be analyzed more easily, but there is still opportunity for mistakes or inefficiencies when the subjects perform the task. The same task was used successfully in our previous study of behavioral patterns [27]. In the current study, subjects did not fully prepare a cup of coffee

but only started the coffee pot brewing so that they did not handle hot liquids, discussed more in section 4.3. While the particular task we used, coffee making, was selected for reasons just given, in the end it is simply a common functional task that serves as a proxy for everyday activity performance. We expect that studying a range of similar tasks would be necessary before our methodology would be put into place in individuals' households, so that an individual who was, say, a tea- rather than coffee-drinker could still benefit from the approach.

4.2 Selection of Technology

Radio Frequency Identification (RFID) technology has been used successfully to study object-use interactions in several activity recognition projects (as discussed in section 3.2). RFID uses tags placed throughout an environment, along with readers that detect nearby tags. An important advantage of this technology is that one can use passive RFID tags which require no power source. Other advantages to using RFID are that tags are available in a small form factor (approximately the size of a postage stamp) and that they are inexpensive (less than \$0.20).

Following earlier work in automated activity recognition, we had each subject wear an RFID reader on the wrist, and we thereby recognized the objects with which the subject was interacting. Specifically, our subjects wore the Intel iBracelet, with a range of about 10cm, that is depicted in Figure 1 [28].



Fig. 1. The Intel iBracelet RFID Reader

This design is beneficial for privacy concerns as well: a subject may take off the bracelet to avoid observation. Likewise, other individuals will not confuse the system as long as they don't wear the bracelet. Disadvantages include the fact that the bracelet is somewhat bulky and has a relatively short battery life (about three hours). If our methodology were used in the home environment, however, the subject would only need to wear the bracelet when making coffee, and could remove it for the rest of the day.

4.3 Experimental Setup

Sixteen subjects with traumatic brain injuries and full neuropsychological evaluations were recruited to participate in the study. For each trial, the subject was to start a pot of coffee—putting in water, a filter and ground coffee and turning the coffee maker’s power on. Subjects were each asked to perform five trials on five different days (13 completed at least three trials and 10 completed all five trials).

The subjects performed the trials in a kitchen at the medical center where they were receiving care for their cognitive impairment. The coffee pot and all supplies were placed on a counter in the kitchen, next to a sink for water. Subjects were asked if they knew how to make coffee and given basic instructions if they did not. No physical demonstrations were given. If subjects asked how much material to put in, they were told to use enough for six cups of coffee (about half the capacity of the coffee pot).

The material that was set out included the coffee pot and carafe, a canister of ground coffee, a bag of filters, a mug and a spoon. Twelve tags were used: four on the coffee pot, four on the canister of ground coffee, and one each on the other objects. Multiple tags are needed for some objects to reliably detect interaction due to the range of the iBracelet (the shorter range is desirable to avoid a higher rate of false positives).

The experimental setup was influenced by the fact that the subjects had cognitive impairments and were performing the task within the clinic. We placed the supplies on the counter, rather than away in cabinets, to make the task easier for the subjects to complete in order to avoid causing frustration by having subjects searching in an unfamiliar kitchen if they forgot where a supply was located. This should not be necessary when observing subjects in a home environment.¹

5 Automatic Assessments

The sensor data collected in each trial consists of a series of time-stamped interactions with RFID tags, a sample of which is shown in Figure 2. From the collected sensor data, we computed four features that we hypothesized might correlate with the subjects’ cognitive impairments. The features are increasingly representative of the task, ranging from very simple—how long does it take the subject to complete the trial—to much more detailed—how “far off” is the subject’s behavior from a correct instance of task performance.

¹ Out of an abundance of caution and on the advice of the clinic staff, we also did not have subjects pour out a cup of coffee once the pot had brewed. This was to ensure that the individuals would not be handling hot liquids and decrease the potential of injuring patients at the medical facility. This should not be a barrier to using a similar system in-home since we expect that many cognitively impaired individuals regularly make coffee and, anecdotally, several participants in our study noted that they regularly made coffee at home (the percentage of participants who make coffee regularly is unknown since that was not part of the formal interview).

Time Stamp	Tag Detected
1200503935	Carafe
1200503935	Filters
1200503936	Filters
1200503939	Ground Coffee 3
1200503956	Ground Coffee 1
1200503989	Coffee Maker Lid 3

Fig. 2. Stream of time-stamped interactions from a portion of a trial. When multiple tags are placed on one object, a number is given indicating which tag has been detected.

5.1 Trial Duration

Our first hypothesis was that a more severely impaired individual might take longer to prepare the pot of coffee than a less impaired individual, as a result of confusion, mistakes, or simply performing steps more slowly. Therefore, the first feature we computed is the duration of the trial: how long it takes the subject to complete the task.

Given a trial with n detected interactions, we define this feature using the following formula: $TrialDuration(t) = EndTime_n - StartTime_1$ where $StartTime_i$ and $EndTime_i$ indicate the start and end times of the i^{th} action in the temporal sequence of trial t . That is, the feature is simply measured as the time between the first interaction that is detected and the last.

5.2 Action Gaps

Note that the trial duration feature is extremely simple and has very limited representational power: it would not distinguish between two people who are behaving in very different ways, provided only that the total amount of time for each trial was the same. We next moved to a somewhat more representational feature, which is based on the hypothesis that more severely impaired individuals might have more periods during which they were not interacting with any objects, on the assumption that during those periods they are considering what step to take next. The second feature measures these periods of inactivity during the trial which we call Action Gaps. We define the number of Action Gaps with length g of trial t :

$$ActionGaps_g(t) = \sum_{i=1}^{n-1} \begin{cases} 1, & \text{if } StartTime_{i+1} - EndTime_i > g \\ 0, & \text{otherwise} \end{cases}$$

We examine the number of brief action gaps using $g = 3$ seconds and the number of longer action gaps using $g = 10$ seconds.

5.3 Object Misuse

We next moved to a feature that accounts for the specific objects used in task performance. One way of determining whether someone is being effective in

carrying out a task is to examine the number of times he or she interacts with each object used in the task. We thus hypothesized that failure to interact with a required object—e.g. to “touch” the coffee filters—indicates a problem, as does an excessive number of interactions. For the simple task of making coffee, we manually determined a reasonable range of interactions with each object, shown in Table I. The filters, for example, may be used once or twice—once to open the bag of filters and perhaps again if the user closes the bag in a separate interaction (remember that the tag is on the bag of filters, not individual filters themselves). Note that we do not state a maximum number of accepted interactions with the Ground Coffee or the Mug or Carafe (to get water) because these are difficult to define—unlike closing the lid which is one distinct interaction, putting ground coffee in the coffee pot may involve touching the ground coffee multiple times to get several scoops and filling the water may require using the mug multiple times to fill the coffee pot. The Spoon is not included in this feature because it was rarely detected—it would also be difficult to use since it is not required but, like the Grounds, may be used multiple times.

For each trial, we then counted the number of times the subject interacted with each object b ($touch_b$) and determined whether that number was outside the accepted range and, if so, how far outside the range it was.²

$$ObjectMisuse(t) = \sum_{b \in Objects} \begin{cases} 0, & min_b \leq touch_b \leq max_b \\ min_b - touch_b, & touch_b < min_b \\ touch_b - max_b, & touch_b > max_b \end{cases}$$

Table 1. Number of Accepted Interactions for each Object

Object b	min_b	max_b
Lid	2	2
Ground Coffee	1	∞
Filters	1	2
Mug or Carafe (Getting Water)	1	∞
Power Switch	1	1

5.4 Edit Distance

Our final approach to automatically measuring performance moves even further in the direction of matching the subject’s performance to an explicit model of correct performance. With this approach, we begin with a representation of how to make coffee—a “plan” for the task. The plan we used in our analysis

² We also investigated a few variations of the Object Misuse metric, to address the possibility that touching an object too many times could have a disproportionately large impact compared with touching too few times. These variations were approximately as successful as the basic metric here; because the variations and results did not appear to be interesting, they are not presented in this paper.

is a partial order over object interaction, depicted in Figure 3, with “Water” indicating using the carafe, mug, or both to get water from the sink and put it into the coffee maker. Note that the use of the partial order allows us to score as “correct” alternative task executions that are reasonable: we score as correct both executions in which water is added before the filter and ground coffee and those in which those actions are reversed. However, we judge to be incorrect executions in which the power switch is turned on before the filters are used.

We then further constrain our plan for correct executions to those in which object interactions are not interleaved and using filters is followed directly with using ground coffee. These two criteria are added for the same reason: for a basic task like making coffee, we hypothesize that it is more likely that a mistake occurred than that the individual chose to interleave actions (like getting ground coffee, then water, and then ground coffee again). Using filters and ground coffee are kept together because we view them as really one general action: putting ground coffee in the coffee maker.³



Fig. 3. Partially Ordered Plan of Object Interaction for Making Coffee

Although we manually created the plan to represent making coffee, other research on activity recognition has addressed the question of automatically constructing plans for everyday activities by mining the web for descriptions of these activities [29]. Such an approach could be adopted to extend our work.

Once we have a plan that models correct task executions in terms of object interactions, we next have to define a measure of deviations from that plan. For that, we adopted the notion of edit distance, which is frequently used in the literature on natural language processing [30], but which has also been used in prior work on activity recognition [25]. More specifically, we make use of the Levenshtein distance which allows the insertion, deletion, or substitution of a character [31]. We compute the distance between the sequence of observed object interactions and each of the legal executions of the plan for the task.

³ The assumptions we have made in our model may be too constraining—perhaps many unimpaired individuals do interleave using filters and grounds with getting water, for example. This suggests a further elaboration, in which the plans are probabilistic—with the probabilities representing the plausibility of certain sequences being performed. This elaboration, however, is outside the scope of this project.

Note that to compute the edit distance, we merge consecutive interactions with the same object (for example, multiple usages of the ground coffee are just shown once as long as no other objects are used in between). We then define $EditDistance(t)$:

$$EditDistance(t) = \min_{e \in LegalExecutions}(Distance(t, e))$$

With our very simple plan, there are only two legal executions: the one in which placement of the filters and the ground coffee precedes the filling of the water canister, and the one in which these occur in the reverse order. Hence Edit Distance is easy to compute, involving determination of just two distances.

The edit distance is intended to provide a fairly fine-grained measure of the relationship between the “correct” task performance, at least as modeled in our plan, and the subject’s actual performance.

6 Neuropsychological Assessment

Neuropsychological impairments are assessed with a battery of tests that sample a broad range of cognitive domains. Many of these tests assess general functioning, such as intellectual ability. Others are very specific, having been chosen because they are known to be associated with functioning that is mediated by a specific brain locus (e.g., left versus right hemisphere, anterior or lateral frontal lobe, specific sub-regions of the areas that mediate expressive or receptive language), or because they provide critical information about a cognitive domain that is central to performance of everyday activities (e.g., attentional shifting). The measures employed for this type of assessment are meticulously normed, often in the context of multiple samples, such that statements can often be made about a patient’s performance relative to the population at large, to specific cohorts (e.g., those of same gender and similar age or education), or relative to specific clinical comparison groups (e.g., is the profile most consistent with a cerebro-vascular accident, dementia, or depression) [32][33][34]. The neuropsychological assessments we used are given in Table 4 in the appendix.

We obtained the results of neuropsychological tests from the patient records of our 16 subjects to use as ground truth. We then computed the correlations of our computed features with the individual neuropsychological assessments listed, using an individual’s average value over the five trials for each computed feature and applying one-tailed non-parametric evaluation. In addition to the individual neuropsychological assessments, we applied principal component analysis (PCA) to the complete set of neuropsychological assessment data for the subjects in order to examine how well our computed features correlate with larger trends in the assessment data. PCA is a standard statistical technique that finds linearly independent components that explain as much variance in the data as possible. Each component is a linear combination of the assessments where the sum of the squares of the component coefficients is one. The first principal component is the linear combination that has the largest possible variance; the second principal

component is the linear combination that has the largest possible variance and is uncorrelated with the first principal component; the third is uncorrelated with either of the first two components, and so on. To perform the PCA, some of the summary assessments in Table 4 were replaced with their component scores, a standard statistical practice.

The first principal component of the neuropsychological data we used accounts for 26.4% of the total variance in the data and the top five principal components together account for 72.0% of the total variance. Table 5 in the appendix shows the first five principal components and the variance explained by each component. Table 4 indicates which factors, if any, each assessment (or any subtest of that assessment) has a loading of 0.6 or higher.

After computing the principal components, the domain expert on our team (Kirsch) interpreted them. The first principal component includes a diverse set of measures of general intelligence. It appears to be a good proxy for general neuropsychological integrity, including measures of intellectual functioning, verbal and nonverbal reasoning, memory, and complex attention. The interpretation of the lower-order components is less clear, although the second could be seen as a measure of general motor integrity; the third as representing verbal memory and concept formation; the fourth, the ability to retain verbal information over time; and the fifth, strategy formation and modification.

7 Results

7.1 Assessing Neuropsychological Integrity

Recall that the main question we ask in this study is whether we can assess a patient's cognitive status by observing performance of an everyday activity using wireless sensor networks. Our main result is quite promising: we found a statistically significant correlation ($p < 0.01$) between the Edit Distance feature and the first principal component of the neuropsychological assessments, which, as just described, can serve as a proxy for overall generalized neuropsychological integrity. Importantly, we did not find such a correlation with any of the simpler features (Trial Duration, Action Gaps, or Object Misuse). The ability to predict neuropsychological integrity, at least within the scope of this experiment and in particular for the population of TBI patients involved, is a strong indication that it is possible to conduct the types of automatic assessments that motivate this work. Figure 4 shows the plot of Edit Distance and the first principal component, with the regression line.

7.2 Assessing Other Metrics of Impairment

Although general neuropsychological integrity is a very important metric, it is also interesting to see how our features assess other metrics of cognitive impairments. The reason for doing this is based on domain practice—in addition to a concern with overall neuropsychological integrity, it is often important for a rehabilitation team to understand different aspects of a patient's impairment: does it

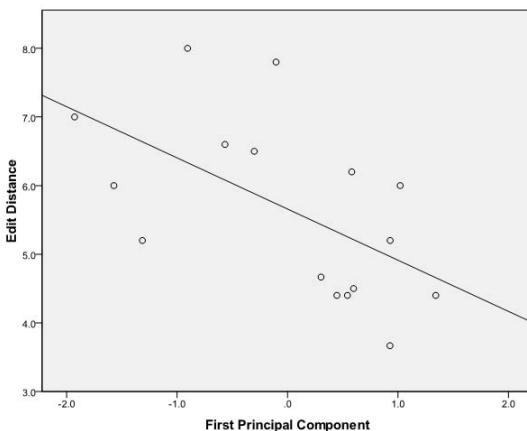


Fig. 4. Plot and Regression Line of Edit Distance with the First Principal Component of the Neuropsychological Assessments

involve memory impairment? Problems with focus of attention? Decreased motor coordination? There is a potential for these to be blurred in a single measure of overall integrity, important though that summary measure is. For instance, an individual whose impairment involves decreased motor coordination or processing speed may have unimpaired executive function, and thus still be able to follow a “plan” for making coffee successfully, but perform the task more slowly. An increased Total Duration might help tease out the types of cognitive difficulties facing this patient. To address this, we also look at the next four (the second through fifth) principal components as well as the twenty-nine individual assessments.

Additional statistical analysis such as a Bonferroni correction is required to state that a correlation between two variables exists with statistical significance. However, because of the large number of assessments and features, achieving statistical significance at this strict level would require the collection of data from a huge number of subjects—many more than were in the scope of this project. Nonetheless, our results on the individual tests in this section are important as exploratory data analysis and as a foundation for further research in the area. Although we recognize that correlations at the variable level are of questionable significance because of the number of analyses performed, we are nevertheless presenting these findings because the coherence of the relationships and the number of associations we found provide important direction for future research. Additionally, while only suggestive, the variable level correlations provide tentative guidance in regard to further refinement of “markers” that clinicians can use when attempting to make a determination of the mechanisms for a patient’s failure. Analysis of mechanisms (e.g., decreased processing speed vs. executive functioning) may then lead, in turn, to choices on the clinician’s

part about interventions that are specifically targeted to the underlying impaired cognitive mechanism.

Beyond this, it is noteworthy that we saw more correlations than would have been expected by chance (22 actual correlations for the individual assessments versus 14.5 expected by chance), especially when looking at what would be strict p-values if the Bonferroni correction were not required ($p < .01$: 8 actual correlations versus 2.9 by chance). Moreover, many of the identified correlations “make sense” from a neuropsychological standpoint, in a manner similar to the example in the previous paragraph. The results from correlations with principal components will also be presented, although the number of correlations with the second through fifth principal components (2) is what was expected by chance.⁴

Edit Distance. The Edit Distance feature achieved the best results with the individual evaluations as well, having a suggestive correlation with the fourth principal component as well as with 7 of the 29 (24%) neuropsychological assessments.⁵ Recall that the fourth principal component appeared to represent the ability to retain verbal information over time. The correlations with individual evaluations are predominantly and compellingly with memory features; we speculate that they could also be said to measure the integrity of the left-cerebral hemisphere and the capacity to engage in sequential and logical thinking. Generally the assessments that have suggestive correlations with Edit Distance are also factors with a high loading in the principal components with which Edit Distance is correlated but the slightly weaker interpretation is likely due to the less efficient analysis of individual assessments. Table 2 summarizes the Edit Distance results and compares them to the other features. Table 3 shows more detail, giving the assessments with which Edit Distance had a suggestive correlation.

Table 2. Summary of Results from each Feature

Feature	Correlations with Principal Components	# Suggestive Correlations with Individual Features
Edit Distance	1st ($p < 0.01$) Suggestive: 4th	7
Total Duration	-	6
Action Gaps ($\geq 3s$)	-	5
Object Misuse	-	3

⁴ These numbers include the results from the additional five variations of Object Misuse noted in Section 5.3 although those results are not presented here.

⁵ We identify a suggestive correlation whenever there would be a statistically significant correlation if the Bonferroni correction were not needed. Because it is necessary, these correlations are not significant but are still of interest for their value in guiding future studies.

Table 3. Suggestive Correlations between Neuropsychological Assessments and Computed Features (Assessments with no suggestive correlations are not shown)

Assessment	Computed Feature			
	Edit Distance	Total Duration	Action Gaps $\geq 3s$	Object Misuse
WAIS III Processing Speed		* #		
WMS-R Visual Reproduction II	*	*		*
CVLT II Total	*			
CVLT II Long Delay Free Recall	*	*		*
CVLT II Discriminability	*			*
Trails B			* #	
Animals	*	*	*	
WRAT 4 Reading			* #	
TPT Memory	*			
Finger Tapping - Dominant	*			
Finger Tapping - Non-Dominant		* #	* #	
GPB - Non-Dominant		* #	* #	

* indicates a suggestive correlation.

notes additional coverage: a metric not also correlated with Edit Distance.

Total Duration and Action Gaps. Total Duration and Action Gaps also proved to be promising features. Though neither had suggestive correlations with any of the principal components, they did with a number of neuropsychological assessments. Total Duration had a suggestive correlation with 6 of the 29 (21%) neuropsychological assessments. Similarly, Action Gaps of 3 seconds or greater suggestively correlated with 5 (17%) of the neuropsychological assessments. These results are less coherent from a neuropsychological perspective than the Edit Distance results but the correlation between processing speed and Total Duration is very logical. And while the results are not as good as the Edit Distance results, they are still valuable: between the three features presented thus far, there are suggestive correlations with over 12 of the 29 neuropsychological tests (40%), including 5 that did not have suggestive correlations with Edit Distance. We also tested Action Gaps of 10 seconds or greater but this only had a suggestive correlation with one metric (GPB - Non-Dominant which also correlated with two other features); we hypothesize that the poor result for this feature is due to the low frequency of gaps that long.

Object Misuse. The results from the Object Misuse feature were the least successful—as shown in Table 3 the feature had fewer suggestive correlations than Edit Distance, Total Duration or Action Gaps of 3 Seconds, and none with assessments that were not also correlated with Edit Distance.

8 Discussion and Conclusion

We have presented an approach to using RFID-based sensing of individuals as they perform a simple task, with the aim of assessing their level of cognitive impairment. We presented four features, with increasingly representational power, that can be computed from the collected sensor data, and evaluated them using the results of the subjects' performance on standard neuropsychological assessments as well as with the principal components of those assessments. The most knowledge-rich feature we computed, Edit Distance, had a statistically significant correlation with the meaningful first principal component, a measure of general neuropsychological integrity. We also presented the results of exploratory analysis of the correlations between the four types of features and the individual assessments; these results are helpful to guide future research into other metrics of impairment without the need for a massive amount of data collection.

There are many practical concerns for the in-home implementation of a system that could automatically assess impairments. Compliance with the system is important since the user must wear the bracelet and complete the task to be assessed; individuals at risk for or developing an impairment may be particularly forgetful about doing this. Other sensor modalities, such as accelerometers placed on the objects, motion detectors, or current or water-flow sensors might be considered which do not have this drawback. On the other hand, the privacy implications of observing individuals in a home environment are important to address and we feel these may be somewhat alleviated by using a system which can clearly be prevented from observing an individual's behavior (by taking the bracelet off).

A great deal of future work remains, including collecting additional data and performing further analysis to investigate the suggestive individual correlations identified in this study. Additionally, further study is needed to examine whether these or similar techniques can successfully differentiate impaired from unimpaired subjects. Observation of other kinds of impairments (particularly dementia) and longitudinal studies of individuals at risk for cognitive impairments are necessary to understand the ability of these techniques to detect the onset of impairment and potentially to develop new techniques to observe change in an individual's performance over time. There are a number of ways in which the scope of the research can be expanded, particularly applying these assessment techniques to other activities beyond coffee making and using them in a home environment. Lastly, studying other types of clinical assessments (such as speech and occupational therapy) as well as development of other computed features, particularly those that might correlate with different assessments from the computed features presented here, are also areas for future work.

Acknowledgments. We thank Intel Research Seattle for the iBracelet RFID reader used in this study. We also thank the staff of MedRehab at the University of Michigan, particularly Erin Spirl, Michelle Shenton and Pippin Gilbert, for their help with subject recruitment and data collection.

References

1. Holsinger, T., Deveau, J., Boustani, M., Williams Jr., J.W.: Does this patient have dementia? *JAMA* 297, 2391–2404 (2007)
2. Wilson, D., Consolvo, S., Fishkin, K., Philipose, M.: In-home assessment of the activities of daily living of the elderly. In: Extended Abstracts of CHI 2005: Workshops - HCI Challenges in Health Assessment, April 2005, vol. 2130 (2005)
3. Sosin, D.M., Snizek, J.E., Thurman, D.J.: Incidence of mild and moderate brain injury in the united states. *Brain Injury* 10, 47–54 (1996)
4. Zoroya, G.: Scientists: Brain injuries from war worse than thought. *USA Today* (September 2007)
5. Okie, S.: Traumatic brain injury in the war zone. *New England Journal of Medicine* 352, 2043–2047 (2005)
6. Bagley, L.J., McGowan, J.C., Grossman, R.I., Sinson, G., Kotapka, M., Lexa, F.J., Berlin, J.A., McIntosh, T.K.: Magnetization transfer imaging of traumatic brain injury. *Journal of Magnetic Resonance Imaging* 11, 1–8 (2000)
7. United States Census Bureau: International data base (July 2007)
8. United Nations Population Division: World population prospects (July 2007)
9. Ferri, C.P., Prince, M., Brayne, C., Brodaty, H., Fratiglioni, L., Ganguli, M., Hall, K., Hasegawa, K., Hendrie, H., Huang, Y., Jorm, A., Mathers, C., Menezes, P.R., Rimmer, E., Scazufca, M.: For Alzheimer's Disease International: Global prevalence of dementia: a delphi consensus study. *The Lancet* 366, 2112–2117 (2006)
10. Sloane, P.D., Zimmerman, S., Suchindran, C., Reed, P., Wang, L., Boustani, M., Sudha, S.: The public health impact of alzheimer's disease, 2000–2050: Potential implication of treatment advances. *Annual Review of Public Health* 23, 213–231 (2002)
11. Jimison, H.B., Pavel, M., McKenna, J.: Unobtrusive computer monitoring of sensory-motor function. In: Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, September 2005, pp. 5431–5434 (2005)
12. Jimison, H.B., Pavel, M., McKenna, J., Pavel, J.: Unobtrusive monitoring of computer interactions to detect cognitive status in elders. *IEEE Transactions on Information Technology in Biomedicine* 8(3), 248–252 (2004)
13. Jimison, H.B., Pavel, M., Le, T.: Home-based cognitive monitoring using embedded measures of verbal fluency in a computer word game. In: 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2008)
14. Pavel, M., Adami, A., Morris, M., Lundell, J., Hayes, T.L., Jimison, H., Kaye, J.A.: Mobility assessment using event-related responses. In: 1st Transdisciplinary Conference on Distributed Diagnosis and Home Healthcare, pp. 71–74 (2006)
15. Pavel, M., Hayes, T.L., Adami, A., Jimison, H., Kaye, J.: Unobtrusive assessment of mobility. In: 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 6277–6280 (2006)
16. Glascock, A., Kutzik, D.: Behavioral telemedicine: A new approach to the continuous nonintrusive monitoring of activities of daily living. *Telemedicine Journal* 6 (2000)
17. Barger, T.S., Brown, D.E., Alwan, M.: Health-status monitoring through analysis of behavioral patterns. *IEEE Transactions on Systems, Man and Cybernetics, Part A* 35, 22–27 (2005)
18. Hoey, J., von Bertoldi, A., Poupart, P., Mihailidis, A.: Assisting persons with dementia during handwashing using a partially observable markov decision process. In: Proceedings of the 5th International Conference on Computer Vision Systems, ICVS 2007 (2007)

19. Albinali, F., Goodwin, M., Intille, S.: Recognizing stereotypical motor movements in the laboratory and classroom: A case study with children on the autism spectrum. In: Ubicomp (2009)
20. Westeyn, T.L., Kientz, J.A., Starner, T.E., Abowd, G.D.: Designing toys with automatic play characterization for supporting the assessment of a child's development. In: Workshop on Designing for Children with Special Needs at the Seventh Conference on Interaction Design for Children, IDC (2008)
21. Ben-Ari, J., Wang, Z., Pandit, P., Rajaram, S.: Human activity recognition using multidimensional indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(8), 1091–1104 (2002)
22. Liao, L., Fox, D., Kautz, H.: Location-based activity recognition. In: Weiss, Y., Schölkopf, B., Platt, J. (eds.) *Advances in Neural Information Processing Systems* 18, pp. 787–794. MIT Press, Cambridge (2006)
23. Pentney, W., Popescu, A.M., Wang, S., Kautz, H.A., Philipose, M.: Sensor-based understanding of daily life via large-scale use of common sense. AAAI, Menlo Park (2006)
24. Logan, B., Healey, J., Philipose, M., Tapia, E.M., Intille, S.: A long-term evaluation of sensing modalities for activity recognition. In: Krumm, J., Abowd, G.D., Seneviratne, A., Strang, T. (eds.) *UbiComp 2007*. LNCS, vol. 4717, pp. 483–500. Springer, Heidelberg (2007)
25. Patterson, D.J., Fox, D., Kautz, H., Philipose, M.: Fine-grained activity recognition by aggregating abstract object usage. In: ISWC 2005: Proceedings of the Ninth IEEE International Symposium on Wearable Computers, Washington, DC, USA, pp. 44–51. IEEE Computer Society, Los Alamitos (2005)
26. Lester, J., Choudhury, T., Kern, N., Borriello, G., Hannaford, B.: A hybrid discriminative/generative approach for modeling human activities. In: Proceedings of International Joint Conference on Artificial Intelligence (July 2005)
27. Hodges, M.R., Pollack, M.E.: An ‘object-use fingerprint’: The use of electronic sensors for human identification. In: Krumm, J., Abowd, G.D., Seneviratne, A., Strang, T. (eds.) *UbiComp 2007*. LNCS, vol. 4717, pp. 289–303. Springer, Heidelberg (2007)
28. Smith, J.R., Fishkin, K.P., Jiang, B., Mamishev, A., Philipose, M., Rea, A.D., Roy, S., Sundara-Rajan, K.: Rfid-based techniques for human-activity detection. *Commun. ACM* 48(9), 39–44 (2005)
29. Wyatt, D., Philipose, M., Choudhury, T.: Unsupervised activity recognition using automatically mined common sense. In: Proceedings of AAAI 2005 (July 2005)
30. Kukich, K.: Techniques for automatically correcting words in text. *ACM Computing Surveys* 24(4), 377–439 (1992)
31. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10, 707–710 (1966)
32. Smith, G.E., Ivnik, R.J., Lucas, J.: Assessment techniques: Tests, test batteries, norms and methodological approaches. In: Morgan, J.E., Ricker, J.H. (eds.) *Textbook of Clinical Neuropsychology*, pp. 38–58. Taylor & Francis, New York (2008)
33. Grant, I., Adams, K.M.: *Neuropsychological Assessment of Neuropsychiatric Disorders*. Oxford University Press, Oxford (2008)
34. Lezak, M.D.: *Neuropsychological Assessment*, 4th edn. Oxford University Press, Oxford (2004)

Appendix: Neuropsychological Assessments Tables

Table 4 lists the individual neuropsychological assessments that we use as ground truth to measure the severity of a subject's impairment, with any principal components for which the assessment had a loading of 0.6 or higher. Table 5 gives the percentage of variation attributed to each principal component.

Table 4. List of Standard Neuropsychological Assessments Used. Parentheses Indicate Principal Components in Which the Assessment has a Loading of 0.6 or More.

Wechsler Adult Intelligence Scale (WAIS) III Verbal Comprehension (1)
WAIS III Perceptual Reasoning (1)
WAIS III Working Memory (1)
WAIS III Processing Speed (1,3)
Wechsley Memory Scale-Revised (WMS-R) Logical Memory I (3)
WMS-R Logical Memory II (3)
WMS-R Visual Reproduction I
WMS-R Visual Reproduction II
California Verbal Learning Test II (CVLT II) Total (1)
CVLT II Long Delay Free Recall (4)
CVLT II Recall Discriminability (4)
Trails A
Trails B (2,5)
Booklet Category Test (BCT) Error Total
Wisconsin Card Sorting Test (WCST) Concepts (3)
WCST Perseverative Errors (5)
Controlled Oral Word Association Test (COWAT-FAS) Total (1)
Animals (1)
Wide Range Achievement Test (WRAT) 4 Reading (1)
WRAT 4 Mathematics (5)
Peabody Individual Achievement Test-Revised (PIAT-R) Reading Comprehension
Peabody Picture Vocabulary Test-Revised (PPVT-R) (1)
Tactual Performance Test (TPT) Total (2)
TPT Memory (1)
TPT Location (2)
Finger Tapping Test - Dominant
Finger Tapping Test - Non-Dominant
Grooved Pegboard (GPB) - Dominant (2)
GPB - Non-Dominant (2)

Table 5. Principal Component Analysis of Neuropsychological Assessments

Component	% of Variance	Cumulative %
1	26.4	26.4
2	15.0	41.5
3	12.8	54.2
4	9.0	63.3
5	8.7	72.0

Further into the Wild: Running Worldwide Trials of Mobile Systems

Donald McMillan, Alistair Morrison, Owain Brown,
Malcolm Hall, and Matthew Chalmers

Department of Computing Science, University of Glasgow, UK
`{donny,morrisaj,owain,mh,matthew}@dcs.gla.ac.uk`

Abstract. Many studies of ubiquitous computing systems involve deploying a system to a group of users who will be studied through direct observation, interviews and the gathering of system log data. However, such studies are often limited in the number of participants and duration of the trial, particularly if the researchers are providing the participants with hardware. Apple's App Store and similar application repositories have become popular with smartphone users, yet few ubiquitous computing studies have yet utilised these distribution mechanisms. We describe our experiences of running a very large scale trial where such a distribution model is used to recruit thousands of users for a mobile system trial that can be run continuously with no constrained end date. We explain how we conducted such a trial, covering issues such as data logging and interviewing users based in several different continents. Benefits and potential shortcomings of running a trial in this way are discussed and we offer guidance on ways to help manage a large and disparate user-base using in-application feedback measures and web-based social networking applications. We describe how, through these methods, we were able to further the development of a piece of ubiquitous computing software through user-informed design on a mass scale.

Keywords: Evaluation Techniques, Large Scale Deployment, Trial Methods.

1 Introduction

It is often considered beneficial to conduct trials of ubiquitous computing (ubicomp) systems ‘in the wild’ i.e. in uncontrolled contexts and environments that are typical of everyday use of many modern technologies [1]. In contrast to the lab-based environment of more traditional usability-style studies, it has been argued that experiments carried out *in situ* can help evaluators gain insight into how people fit systems into their existing practices and contexts of use, and how people change their contexts and practices to accommodate or take advantage of new systems. While this approach has its benefits, the staging of ubicomp system trials may give rise to a number of practical issues that inhibit evaluators’ ability to draw substantive conclusions on system use. For example, many trials involve providing each participant with a mobile device on which to run the system under investigation. This in itself can introduce biases into the trial: participants are dealing with a device with which they are not familiar, and there

will likely be a period of acclimatisation during which they might not use the new technology as naturally, or with the same degree of skill, as more experienced users. Merely having to carry around an extra device during a long-term trial might have an effect on some participants—it is likely that they will already be carrying mobile phones and perhaps also cameras, so the obligation to carry around additional hardware might affect participants' perceptions of the system, or they may simply not always carry the trial device around or use it as much as experimenters hope or expect.

Another limiting factor that arises when providing participants with new hardware on which to run a system under investigation is the number of devices that can be supplied, and therefore the number of participants available for the trial. Most research projects have a specific budget for trial hardware, but this rarely stretches to pay for thousands or even hundreds of devices such as smartphones, and so the size of experiment that can be conducted is necessarily limited to a relatively small number, e.g. 10–20. Such hardware may be shared by several experiments in the same project, or in several projects, and this may create pressure to keep trials short so that different experiments can take place.

Furthermore, if participants are supplied with devices by researchers, it is common practice to recruit these users from the researchers' local area. Many university-based research teams will use student volunteers as participants, for example, or other participants who reply to adverts placed around the campus. Although many interesting findings are of course possible from such a user-base, an evaluator could not realistically extrapolate these insights into conclusive statements in a global sense; how a group of university undergraduates adopt a particular technology may not be typical of the wider community in the same urban area, and communities in a different continent may be even more different. So, not only does a local participant set give rise to the dangers of basing findings on a very narrow subset of a technology's potential user-base, it also leaves no possibility for studying cultural differences by comparing many geographically distant groups of users.

A step towards addressing some of these issues is running a trial of a ubicomp software system not on experimenter-supplied devices, but on devices the participants already own and use daily. Only in very recent years have we seen mobile phones that are both numerous enough to afford a large trial as well as advanced enough to support downloading and installation of researcher-supplied software. Market research firm IDC [2] suggests that, at the end of 2009, 15.4% of the mobile phone market consisted of smartphones, an increase from 12.7% in 2008. So, while still not the predominant type of handset, we suggest that smartphones have been adopted into mainstream use. While running a trial solely with smartphone owners may not be selecting a user-base that is representative of the population at large, it is not now using only the most advanced 'early adopters'. By recruiting smartphone owners, we may be able to avoid or reduce some of the issues outlined earlier, in particular the small number of hardware devices that a research project can generally supply and the length of time that a trial can last for.

In this paper we describe our tools and techniques for recruiting smartphone owners for ubicomp trials, deploying systems amongst them, directing questions to users and encouraging social interaction among them. We document our experience of a system's deployment among 8676 active users. A key element of our recruitment and deployment was our use of a public software repository rather than directly supplying

software to trial participants. Although a recent phenomenon, such repositories are a well-established means of deploying software to smartphone users. Apple's App Store has proved to be a very popular and effective means by which iPhone users can access new software, and several other mobile platforms now have similar repositories. Despite their popularity, the potential for such repositories to be used as a distribution mechanism for research prototypes, while having been touched upon in [3], has not yet been explored and documented, and yet several potential benefits of such a mechanism are apparent, e.g. such repositories already offer means for users to browse and find software they are interested in, and so researchers can effectively advertise a system and recruit participants for a trial by putting the system into a repository.

This paper describes our experiences of making a free application available in this way, a mobile multiplayer game called Hungry Yoshi. This is a new version of Feeding Yoshi, a seamful game that we ported to the Apple iPhone and updated. Feeding Yoshi's main trial was described in [4] as a "long-term, wide-area" trial "being played over a week between three different cities" in the UK. We wished to scale up our deployments and trials as part of a project, Contextual Software, that explores system support for collaboration with communities of users in the design and adaptation of software to suit users' varied and changing contexts [5]. Distribution in the App Store style, along with our new tools and infrastructure, allowed for a trial that involved a much larger number of participants than before, who were far more geographically dispersed than we could previously handle, and which lasted longer than any trial we have ever run. At the time of writing, the current trial of the new Yoshi system has been running for five months and has involved thousands of users from all around the world.

The following section describes work related to this and other examples of large-scale trials, as well as outlining the original Yoshi system and trial. This is followed by a description of the re-design of Yoshi for use on the Apple iPhone and wide-scale distribution. Thereafter we describe the processes involved in distributing the game to a global audience, managing a trial involving a large and widely distributed user-base, and involving those users in development of a new system feature. We then discuss some methodological and practical issues before we offer our conclusions.

2 Related Work

Several ubicomp projects have featured data collected from large numbers of people via mass-scale sensing. An example is the Cityware project [6], which collected data from scans of Bluetooth devices detectable by static recording equipment at various locations around a city in order to measure densities and flows of people in particular urban areas, which in turn were to be used in architecturally based models of those areas. In a related theme, abstractions similar to those of the Cityware work but at an even larger scale were shown in [7], which involved the generation of coarse-grained city-scale maps of people's density based on concentrations of mobile phone signals sampled from GSM infrastructure. While this work undoubtedly exhibits great scale, it is different to the area we are investigating in that sensor data is collected and aggregated, rather than data on the use of applications. Also, such techniques do not directly feed into qualitative investigations of social and personal behaviour, a useful combination that we aim to support.

In 2008 Nokia Research Centre released Friend View, a “location-enhanced microblogging application and service” [8] via Nokia’s Beta Labs. This site allows its community members to contribute feedback to in-development and experimental software, but this study reported only on statistical analysis of social network patterns based on anonymised log data representing 80 days’ use by 7000 users. Like Cityware, this serves as an example of many quantitative studies in which potentially interesting analyses were carried out, but no interaction with users is described that would allow analysis to be contextualised with user experience, or determine how users’ opinions, behaviour or systems might change in the light of such analyses.

One of the early landmarks of large-scale deployment of ubicomp applications was *Mogi Mogi*. As reported by Licoppe and Inada [9], this location-based mobile multi-player game was released commercially in Japan, and in 2004 had roughly 1000 active players. Some basic aggregate analyses involved system profiles, e.g. gender and age group, but almost all the presented analysis is based on more ethnographic interviews and observations of ten players who were, apparently, strangers to each other. This method afforded rich detail of the ways that they fit the game into urban contexts and lifestyles, based on months of game play, including occasional social interactions between players.

In 2006, we trialled Feeding Yoshi, running what we called “the first detailed study of a long-term location-based game, going beyond quantitative analysis to offer more qualitative data on the user experience” [4]. The participants consisted of four groups of four people and, as mentioned above, the main study lasted a week. The participants in each group knew each other before the trial, and collaboration and social interaction was observed during the trial. The study drew on participant diaries and interviews, supported by observation and analysis of system logs. Somewhat like the study by Licoppe and Inada, it focused on how players “interwove the game into everyday life” and how wireless network infrastructure was experienced as a ‘seamful’ resource for game design and user interaction.

Observational techniques founded in ethnography may be well suited in principle to studying ubicomp systems, but in practice they are often hampered because keeping up with the activity is difficult, small devices such as mobile phones and PDAs can easily be occluded from view, and people’s use may be intimately related to and influenced by the activity of others far away [10]. Several video cameras may be used to record activities in several locations set within some larger activity, but this brings the practical problem of synchronisation, and how to gain an overview of this material and combine it with other relevant data, such as system logs gathered from the mobile devices. Furthermore, network connectivity may be intermittent or costly enough to hamper attempts to keep in continuous contact with users and their devices, e.g. to stream log data back to evaluators or developers. Consequently, some researchers have explored ‘experience sampling’ methods, in which a questionnaire appears on-screen when the mobile device detects that it is in a context of interest [11]. Carter and Mankoff developed Momento [12], which supports experience sampling, diary studies, capture of photos and sounds, and messaging from evaluators to participants. It uses SMS and MMS to send data between a participant’s mobile device and an evaluator’s desktop client. *Replayer* [13] similarly combined quantitative

and qualitative data in order to offer a more holistic view of systems in use, and to let researchers study users acting in greater numbers, and at larger geographic and time scales, than they can directly observe. In particular, it used spatial properties within quantitative log data so as to make analysis of qualitative data less time-consuming and therefore allow larger trials to be run.

3 Hungry Yoshi

Feeding Yoshi [4] was a mobile multiplayer game for Windows Mobile PDAs. It was re-implemented for the Apple iPhone and renamed Hungry Yoshi. It uses wireless networks infrastructure as part of its design. Players' mobile devices perform regular scans of the WiFi access points visible in the current area, classify each of these access points according to its security settings and display it to the player. Each password-protected access point is deemed to be a creature called a 'yoshi' whereas a network without password protection appears as a 'plantation' growing a certain type of fruit. Yoshis ask players for particular fruit, and players score by picking these fruit from the correct trees and taking them to the yoshis. Yoshis also provide seeds that enable players to grow new fruit in empty plantations. A research objective of the 2006 study of Feeding Yoshi was to establish how players could interweave playing a long-term game with their everyday lives. Four teams of four players were used in the trial, each being paid for taking part, with a competitive element introduced such that the members of the team with the highest combined score received double the standard participation fee.

Hungry Yoshi has some differences to Feeding Yoshi. Perhaps the biggest change is that, with the availability of the iPhone's data connections over cellular networks, the system can generally maintain a globally synchronous game world. In the old game, yoshis and plantations visited and their contents were stored only on players' mobile devices, so two players might visit the same plantation and see it containing different contents. By storing such details on a centralised server, one player can seed a plantation with a fruit type and another can pick the fruit when they grow. A new piece of functionality in the iPhone version of Yoshi is the ability to change pieces of fruit for a small cost. Players are able to insert fruit into a fruit swapper (Figure 1-right) that returns a different type of fruit at random. To use this swapper, players are charged tokens, which can be earned by performing tasks. Section 4.3 explains why this task mechanism was important for helping us interact with the users during the trial.

Another difference from the original trial is that the game no longer has any explicit team element: each player participates as a solo entity. However, the score table is retained as a form of motivation for players, though with the difference that now there is no prize at the end of the trial and indeed no defined end to the playing of the game. Separate score tables are maintained for overall score, score this week and score today, the latter two being used because new players might join in at any time and could be months behind the early users. The table screen is divided into a top section showing the top players, and underneath, a section showing the players around the user's current position.



Fig. 1. The list of nearby yoshis and plantations (*left*), a yoshi screen (*centre*), and the fruit swapper (*right*)

4 Engaging Users Worldwide in Iterative Design

This section describes how Yoshi was evaluated and modified in the course of our trial. It discusses our approach to distribution, management, data gathering, analysis and redesign, coupling them together in a form of iterative design suited to the large scale of our trial. We outline how we interacted with trial participants, how users interacted with each other, and how these interactions fed into a new version of Yoshi so as to begin another design iteration.

4.1 Distribution

Hungry Yoshi was released in early September 2009. At the time of writing it has been publicly available for five months. Distributing software via a public repository means using a mechanism that users are already very comfortable with, again possibly leading to more naturalistic interactions than with a more contrived physical meeting and handover of a device or software. Yoshi appears in the ‘games’ section on the store, and so benefits from recruiting users who deliberately seek out this type of application and who will hopefully therefore be more keen to engage with the game. An unanticipated but welcome benefit to this form of distribution is free advertising outside of the store and beyond our own announcements of the trial, e.g. in interviewing one of Yoshi’s users, we learnt that she first heard of the game in a review in an Italian technology blog. In releasing a research prototype through a public marketplace, we harness some of the enthusiasm of amateur and professional writers who regularly scour the store for new applications to try and discuss.

Figure 2 charts the number of downloads of the game over the time that the game has been available. When the game was first released, and when updated versions are made available, it features near the top of the store’s “most recent” lists, providing a boost in the number of downloads that day.

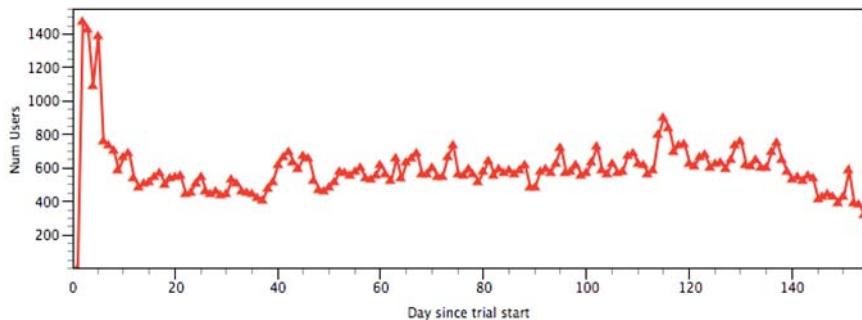


Fig. 2. Number of downloads of Yoshi per day since release

It can be seen that there was a peak of interest in the first few days following the game being launched, after which download figures were around 600 per day. There appears to be a gradual trend upwards, perhaps falling off only in the last month or so. Occasional spikes, such as that at 40 days, correspond to the release of new versions. At the time of writing there have been 137367 downloads in total. This figure includes people updating to new versions of the game; we recorded 94642 unique downloaders. Figure 3 shows the proportion of players of the game each day who are playing it for the first time, as compared to those who have played the game before. It can be seen that by the end of this period, the proportion of returning players is increasing although around 25% of players are playing for the first time each day.

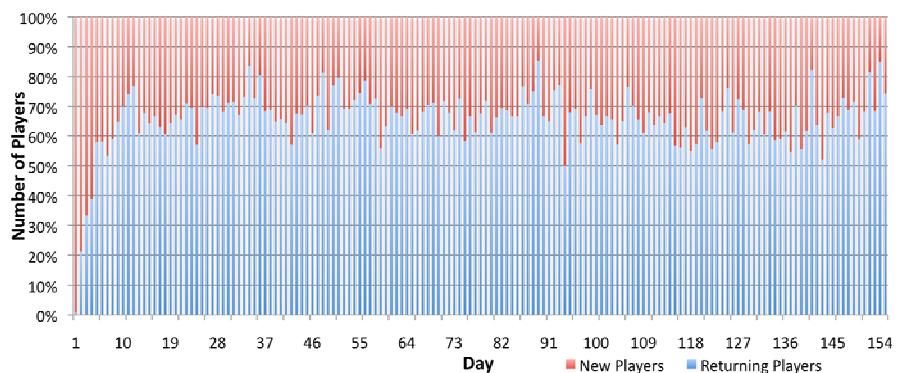


Fig. 3. Proportion of new and returning players per day of the trial

Having installed the game and on opening the Yoshi application for the first time, users are presented with an information page, written in English, French, German and Japanese, that explains that the system is created as part of a research project and that details the various forms of information that will be logged during interaction with the game. The page also states that researchers might contact users to enquire about their use of the software, but that these communications can be ignored and would cease on

request. Only by agreeing that they have read and understood these terms can players proceed into the game. Further to this, we state that log data will be stored securely, that users can opt out at any time, that we will destroy the data logged about them on request and that all the data will be destroyed following the end of the project. To date, no such requests have been received. Links are provided to an email address for the trial organisers, and to a public Web forum where users can either chat amongst themselves or seek clarification on any aspect of the game or research trial from the organisers. The data logging process is described below in Section 4.2. At the time of writing, 24408 out of the 94642 downloaders registered with the game and agreed to be part of the trial. This reduction may be because people were wary of having their data logged in the manner described, were perhaps apprehensive over being contacted by researchers or were deterred by having to register a user account. Of those 24408, many only briefly interacted with the game, but 8676 played for long enough to produce log data that could be studied. Although this represents only around 9% of the total number of downloaders, the number of players is still very large.

Quantitative analysis benefits from having such a large user-base. Having information gathered from thousands of users allows many inferences to be made with a much higher degree of confidence than if an experiment had been run with, for example, the 16 participants we had in 2006. Results of our quantitative analyses are covered in the following section, and we offer some reflections on this scaling up in the later Discussion section.

4.2 Quantitative Analysis

To aid our evaluation of Yoshi, system log data is generated from every trial participant's phone. The system makes use of our SGLog logging framework (described in detail in [5]), which manages data collection on the phone and periodic uploads to a server using the same data connection required to run the game. The data logged includes activities within the game, such as feeding a particular yoshi, and general contextual information. Uploaded data from each user is timestamped and stored on a database on a central server. To protect the privacy of participants, this framework uses TLS to encrypt data sent between phones and the server.

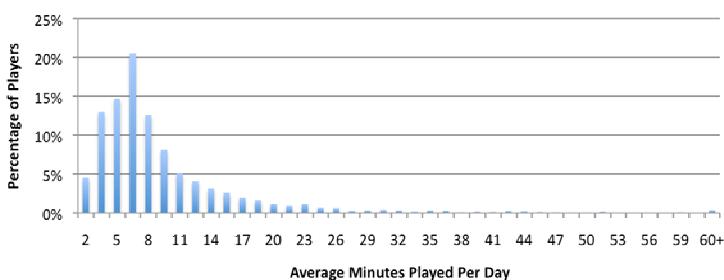


Fig. 4. The distribution of players' average system use per day, with a mode of 6 minutes (20.5% of players)

Figure 4 shows the distribution of the average amount of time each user played the game each day. This time was calculated by looking at timestamped game events registered on the server, rather than simply the times at which the application was running, so times when the device was sitting idle do not contribute to the figures. It can be seen that there is a range of levels of activity, with several players playing for over an hour a day on average.

One player's average daily play was significantly longer than the rest. Over the first two months of the trial, she had an average of more than 2.5 hours of play per day and at the time of writing has played the game for over 200 hours. She is the game's top player, and has been at the top of the overall score table since the early days of the trial, with around double the overall score of the second highest-scoring player. In any trial it is probable that researchers will observe a variety in the level of engagement shown by users. In running an experiment with hundreds or thousands of participants, it is likely that this spread will be wider, and that some of these users will be more enthusiastic. For example, in the original trial of Yoshi [4], the longest time a player spent playing the game in any one day was 2.5 hours, whereas here this figure is almost 7.5 hours.

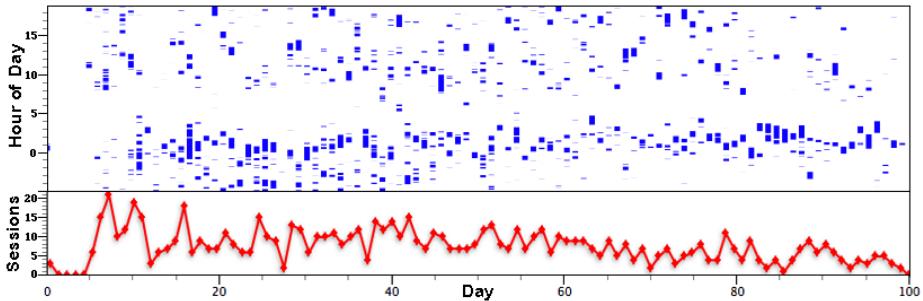


Fig. 5. A visualisation of one of the most active players' use over the first 100 days of the trial. In the upper section, the *x*-axis shows days since the trial began and the *y*-axis shows the 24 hours of the day, with blue shading showing the periods at which the participant was playing the game. The lower section shows the number of 'sessions' per day for the same user, with 'session' meaning a period with less than five minutes between each user action.

Figure 5 shows one of the top-scoring players' activity in greater detail. The number and lengths of lines give a quick impression of the amount of activity this user has engaged in, and the length of these lines shows whether the user favours long sessions or quicker games, squeezing a short burst of play into a spare few minutes. We also see daily patterns, e.g. finding a strongly shaded row in the plot would indicate regular play around that time of day. Quantitatively-based visualisations such as these were useful both in themselves, in letting us see basic patterns of use, but also in feeding into qualitative analysis, e.g. in selecting participants to interact with more directly, and in preparing for such interactions—as described in section 4.3.

4.3 Interacting with Participants

One of the challenges of conducting a worldwide system trial lies in managing interaction with participants: maintaining a presence with participants, harnessing

feedback and supporting qualitative analysis. As we would not be able to regularly meet participants, as one might do in a more standard trial with locally sourced trial subjects, alternative means were sought to keep in contact with our users. We used two mechanisms: tools for communication within Yoshi, and communication via a social networking web site.

In-Game Communication with Users

Rather than have communication with users happen in a way that clashed with the user experience, we built bi-directional communication into the functionality of the Yoshi game. Section 3 introduced the fruit exchange mechanism, which users were charged tokens to use. These tokens were earned by players performing tasks, set by researchers throughout the course of the trial. In this way, we could relay messages to participants, ask specific questions and receive feedback as appropriate.

The tasks set to users in this manner took a number of forms. Simple factual questions such as age, gender and continent of residence were asked, with users selecting answers from drop-down lists. This provided a simple means for us to build up demographic profiles of the user-base. More open-ended questions which allowed free text responses were also set, such as what a player liked about the game, and whether he/she had any suggestions or bugs to report—as we detail later. This system proved to be of particular benefit because the tasks could be dynamically updated in real-time during the course of the trial, and because specific questions could be targeted towards a particular user or set of users in response to some interesting activity we observed in their log data or our interactions with them. The tasks available to a player are downloaded from the server each time the player visits the task list screen. Therefore, although the system is deployed to a worldwide user-base, and we could not access devices to update the software on them, we could alter the questions at any point during the trial. Once edited, the new task set becomes live immediately, thus supporting adaptation of our research interests.

The task and token-earning functionality proved popular with users, with 28442 responses in total. Before the trial, we were unsure whether players would use this feature ‘honestly’ or would provide dummy answers. As no checks were in place, free text answers could be submitted as empty or with a few random characters, and players would still be rewarded tokens by the automated system. However, results proved that users were willing to engage with this feature, providing answers of varying length, but in the main making an attempt to answer in a useful way. As an example, a task asking demographic information from the user was completed 2406 times, with all but 73 being sensible answers to the question. While the tasks themselves were in English, care was taken to ensure that where possible the grammar and vocabulary used fell within the Common European Framework of Reference for Language’s A2 level bounds, a level achievable by most attending public school in westernised countries where English is taught as a second language [14].

Interacting with Participants through Facebook

Although the task system provided a basic communication mechanism between researchers and participants, more powerful external tools were also used in order to facilitate more in-depth dialogues and to support communication between participants

themselves. We elected to use Facebook, a popular online social networking application, as a means of supporting such interactions. Facebook has more than 300 million active users, 50% of whom log on to Facebook in any given day [15], making it an appealing choice of platform for this task. Also of benefit was Facebook Connect, a service with an iPhone API that allows users to verify themselves and log in to third party sites and applications using their Facebook account. On starting Yoshi, players are required to log in to their game account in order to track their score across devices, and to allow multiple people using the same device to have individual accounts. This can be done either through Facebook Connect or by creating a username and password specifically for the game (which we called ‘Lite mode’).

Though we still wanted non-Facebook users to be able to play the game, we sought to encourage users to login through the Facebook Connect method to provide the benefits outlined above. As such, we limited use of the fruit swapper described earlier to only Facebook-logged in users; users logged in via Lite mode were prompted to login to Facebook when attempting to access this functionality. Additionally, we allowed users to post their Yoshi progress to their Facebook Feed (which shared their scores and rankings with all their Facebook contacts). This served both as an enticement to use the Facebook version, and as further user-generated advertising for the game. Of our 8676 users who agreed to the terms and played the game, 6732 elected to use the Facebook login, including 44 of the top 50 scorers.

In addition to providing a login mechanism, we also used content on the Facebook site itself both to provide features for the user and in contacting users to aid in the management of the trial. We created a Facebook application—a series of PHP-based web pages displayed within Facebook—that showed the ranked scores in greater detail and provided statistics on the players’ game play, such as their most visited yoshis. More importantly, Facebook has a set of well-established means of communication both in one-to-one and one-to-many models. For example, as players had provided us with their login IDs, we could send emails to their Facebook accounts, and we set up a forum for players to communicate with each other and discuss potential new ideas.

4.4 Qualitative Analysis

Section 4.2 described quantitative analysis performed on log data. With such data, gaining an in-depth understanding of individual player behaviour is challenging. While we could visualise various aspects of play, this did not necessarily make a player’s motives and reasoning comprehensible. We now describe allied forms of qualitative analysis, centred on interviews that let us explore and clarify issues more adaptively than if, for example, we had used an on-line questionnaire to gather qualitative data. As will be discussed, some of the processes already described such as visualisations and Facebook tools were useful resources for this form of analysis.

Interview Process

The process of interviewing participants worldwide is not quite as straightforward as in a more traditional experiment involving locally based users. Whereas in a traditional setup researchers are likely to have met participants before the trial begins,

perhaps to deploy the system or to explain the trial, we had no direct contact with users at the beginning of our qualitative analysis process. Although all the users had agreed to a series of terms before playing the game, that explained that we might try to contact them, they had also been informed that they could feel free to ignore this communication or to tell us that they were not interested in participating. More positively, having over 8500 users gave us an opportunity to focus on interviewees that we deemed the most relevant to a design issue or potentially significant in terms of use and user experience. For example, we could choose the most active players, i.e. those who had accumulated the most game time, those who had answered a particular in-game question, or those who had a particular pattern of use in their system logs.

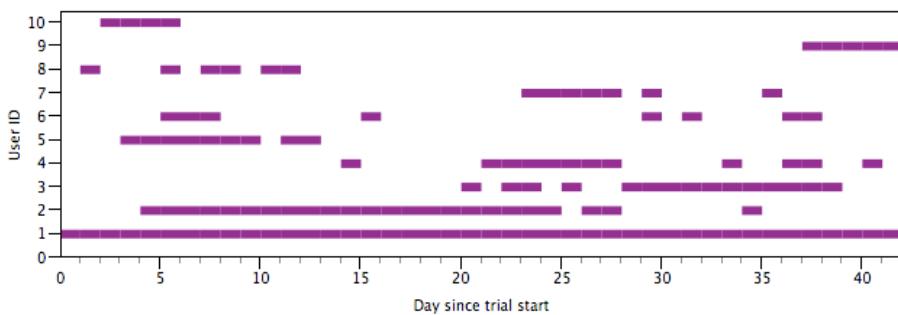


Fig. 6. A snapshot from a tool used to select participants for interview, showing which days each participant used the application. From a chart showing all the trial participants, ten users have been selected who exhibit contrasting patterns of use. This figure shows the filtered set.

Selecting participants for interview began by using the information publicly available to application developers via the Facebook API to first filter our participant database to show only those over the age of 18. Thereafter, we used visualisations of log data to examine participant use over time. For example, Figure 6 has a separate row for each participant, and shades days, shown on the *x*-axis, if the user played the game on that day. We were interested in speaking to a set of users with a diverse amount of engagement, so used the visualisation to select rows with contrasting patterns. Figure 6 shows such a subset of users. As can be seen, there is a wide variety of use, with user 1 playing every day, user 10 playing for a few days near the start of the trial before giving up and user 9 joining the game later than the others, but having played every day since starting until the current time. Simpler methods of selecting interviewees would obviously have been possible, such as choosing the highest-scoring players, but we were interested in using our methods of selection to interview a set of players which showed a broader range of activity in playing the game.

Having identified our interviewee set of choice, Facebook again proved a useful tool in making contact. We emailed users, enquiring whether they were interested in taking part in an interview over VoIP or telephone in exchange for \$25 of iTunes or Amazon vouchers. We interviewed 10 of these players, from 5 different countries and 3 different continents. 5 were male and 5 were female, and they ranged in age from 22 to 38. As the trial progressed, we noted a shift in participants' willingness to be interviewed. In the first months of the trial, requests for interviews were met with

enthusiasm and even pleasure at being given the opportunity to participate. However, the number of users willing to take part in the feedback process seemed to drop as the trial continued. We speculate that this is perhaps a result of the self-selection of participants involved in this type of trial; early adopters willing to download software on release and persevere through initial versions seem to have more interest in the process and greater willingness to participate than those who joined the trial later.

In addition to the visualisation shown in Fig. 6, other tools were useful in interview preparation. The tool seen in Fig. 5, showing activity across different hours of the day, helped to make the interviewer aware of the overall level of engagement the player had shown, as well as raise specific items of interest to ask about, such as if a player seemed to be using the application for five minutes every day at lunch time. Similarly, the answers that the interviewee had submitted to tasks were surveyed before the interview commenced, again to prime any potentially interesting aspects to ask about. Each interview began with another explanation of the trial and of who we are, and typically lasted between 15 and 45 minutes. All were transcribed afterwards.

Findings from Analysis of Interviews

In order to explore our research issue of running trials at a distance, we asked as many questions about the trial mechanism itself as about the application. Unsurprisingly perhaps, many of the same game experience themes arose in interviews as had been reported in the original Yoshi trial. For example, several participants mentioned that awareness of other players' scores, as shown in the table on Facebook and within the game itself, was a strong motivating factor for them. It should be noted that, unlike the first trial of Yoshi, no prizes are awarded to the best players; presenting names on the score table and the ability to share their success with all their friends, players and non-players, via Facebook, proved to be enough of an incentive for many players.

Our use of Facebook also afforded more direct interaction between players. By having full names visible on the scoreboard, and as the game had clear links to Facebook, users appeared to have a 'ticket to talk' to each other. For example, one participant (A) reported seeking out another player (B) on Facebook, to ask about what was perceived as unusual scoring patterns. From seeing B's name, A made assumptions about where B was likely to be based in the world and was confused about the times of day that B appeared to be accumulating points: "*I really couldn't figure out how they could have all those points when I was asleep*". After exchanging a few emails with each other, A discovered that B lived in a different continent. Their email conversation has continued, and they now consider themselves friends.

Turning now to the user trial mechanisms, interviewees were enthusiastic about the range of feedback mechanisms made available to them. In particular, players we interviewed were very positive about the task mechanism, with one saying "*I think it's a pretty good idea that I can answer certain questions for [tasks] so I can give feedback there. Even free-text feedback. And it's really good.*"

This trend is in accord with our analysis of task response rates and the number of sensible answers received. One interviewee specifically addressed having noticed that it was possible to just get tokens from submitting empty responses, but still felt he should give proper answers: "*Sometimes, you scan through, and just try and hit the submit button ... you're just like, gimme these tokens, I wanna get on with it... But most times, I answer honestly, about 98% of the time.*"

This enthusiasm for the task response mechanism extended to being emailed over Facebook to request an interview, with all interviewees responding positively when asked how they felt on being contacted, with one commenting: “*I find it really nice that [you are] contacting me and asking me my opinion. I guess it’s a really nice thing.*” Indeed, at the end of their interviews, two of the ten interviewees actually declined the payment that had earlier been agreed, saying that they were happy to participate. We speculate that this is maybe because we had provided a free game that these players evidently value. Of course, players who do not enjoy the game stop playing it and are not available for logging or interview—thus potentially biasing our ‘sampling’ of users. By targeting users with the task mechanism, Facebook messages and email we were able to quiz those who declined interview requests on their reasons for doing so. The response rate was low but those we did receive fell evenly into categories of general refusal, e.g. ‘I don’t have time.’, and refusal based on perceived lack of language skills, e.g. ‘I don’t speak English.’

Users are playing of their own free will rather than perhaps feeling obligated by having agreed to participate in a system trial, and so their play is more ‘natural’ than those who use the system out of a sense of obligation or for financial benefit. As a result, compared to our experience of earlier trials of other systems, we observed that players had more good will towards ‘giving something back’ than we have observed in more traditional trials.

Although time-consuming to arrange and conduct, these interviews offered valuable insights into player behaviour and their reactions to the trial process and provided a valuable, rich communication channel through which detailed contextual understanding of logged data could be sought.

4.5 Redesign

Given the flexibility of the tools for interacting with users and studying log data, we were able to use the tools to ease the task of redesign. This reflects one of our research goals, which is developing means to quickly and appropriately adapt software to suit the changing contexts and interests of users.

For example, as an answer to the task “What could be improved about Yoshi?”, one user (anonymised here as Helen) commented that plantations were often too full. Helen was invited for interview, and the interviewer then raised this point to obtain further detail. Helen explained that, as plantations auto-generate fruit at a rate of one per hour, they would often be full, which she felt was to the detriment of the game. In particular, Helen described a situation where she would empty a plantation before leaving for work in the morning, and wanted to collect a seed from work to plant when she got home. However, by this time the previously empty plantation would have around 10 pieces of fruit in it again, which would have to be picked first and fed to unwilling yoshis, leading to a points penalty.

Following this interview, the game designers agreed that this was a valid criticism that should be addressed if it reflected a common concern or problem among users. We again used the task mechanism to consult our user-base at large. A question was added as a task in the game, in the form of a vote as to whether to introduce this feature, and exactly what form it should take. We presented three options: (A) leaving the game unchanged, (B) players could burn empty plantations to stop them

re-growing (as suggested by our interviewee) and (C) even full plantations could be burned, which would also destroy all the fruit that had grown. 17% voted in favour of leaving the game as it was, while 29% were keen to see option B and 54% selected option C. The chosen feature was therefore implemented and distributed in a new Yoshi version, thus beginning another iteration in our design process.

On detecting that Helen had installed the new version, we contacted her again to gauge her reaction towards the new feature and she replied positively, agreeing that the version implemented, although not the design she had suggested, was the better of the new options. Around the same time, we included another vote on the new feature, consulting the opinion of the user-base at large after they had had a chance to use it. Users responded with approval, with 94% agreeing that they liked the new feature. This demonstrated to us a significant benefit in this iterative approach of conducting design by engaging with users at both a micro and macro-scale, and letting the results of one feed into the other.

System bug handling was dealt with in the same way. One user was having stability issues that were reported in-game through the task mechanism. Upon contacting the user for more information, the problem was narrowed down to be specific to his model and operating system version combination in areas of high access point saturation. This problem was resolved and the next update to the game was released. Over the five months the software has been live, seven versions have been released to the public.

By having interaction with evaluators integrated into the game dynamic, users are able to report issues directly within the relevant context of use. While these reports are generally brief, they provide a hook back to the context of use they were created in. In this respect, the log data was an invaluable tool for helping the user recall the context of use and therefore the detail and qualitative texture of the problem or suggestion he/she had reported previously. Placing the user at the scene of the problem or suggestion by discussing their location, the game actions they took leading up to the report, and how their pattern of play had evolved to the point where they noticed a problem gave interviewers a valuable means to elicit the detail necessary to pinpoint problems and ground suggestions.

5 Discussion

The tools and techniques described in the previous section let us carry out a relatively normal iterative design process but at an unusually large scale. Methodologically, when we compare our approach to more standard trials, we see both advantages and disadvantages. The large number of users is helpful in statistical terms, but the volume of data can inhibit the move from quantitative aggregates to qualitative detail. While we sometimes used common database query tools to work with the ‘raw’ log data, we found it beneficial to also develop our own visualisations to better understand patterns and detail in the data and to choose where to focus requests for interviews. Compared to more traditional trials that involve local participants who are paid to use our software in a trial, we suggest that our process of ‘recruitment’ led to more realistic conditions in that users were using software that they themselves chose to use—without inducement from us or obligation on their part to keep using it even though they did not want to. However, this advantage has to be weighed against issues such as our inability to gather data from those who dislike the application, and our reduced knowledge of local context and culture.

In practical terms, our methods incurred expenses in terms of development time and interviewer effort. The language skills of the group were put to the test as we created French, German and Japanese internationalisations—giving greater access to those for whom English is not their first language. Initial worries about our ability to interview users with limited English and a first language outwith the skill set of the team proved to be irrelevant, however. The nature of the interview selection process meant those who were not confident in their language skills were less likely to volunteer to be interviewed. Again we note a potential bias: potentially significant interview subjects could decline to be interviewed due to their lack of confidence.

Communicating across time zones can cause delays and sometimes involved the scheduling of out-of-hours interview times in order to fit in to the daily schedules of our users. Taking into account the time differences when considering the rapidity of response from users is another aspect; users generally expect ‘timely’ responses to any messages they send—no matter how many time zones away from the developers they are. We found that taking careful note of the sender’s time of day when a message was created, and addressing their perception of the passage of time until we responded, was important in building relationships with users, e.g. a reply which will not be read until the ‘next day’ in the user’s time zone should be phrased to take into account the user’s likely perception of a slow response.

In our trial we found that relative wealth scales also played a part, with the level of entry to our trial set at having an Apple iPhone—still a relatively expensive item which is not price-normalised to match local incomes. In rough terms, and taking into account countries’ populations, we observed that as the average income of a country decreased so did the density of Yoshi users there. We suggest that this pattern may not appear with software for more widespread, price-normalised mobile phones—potentially leading to a larger proportion of users in countries with lower average incomes taking part in trials. Similarly, although the trial software was developed on the latest iPhone hardware, firmware and OS, care was taken to ensure that the game was backwards compatible with older versions to try to maximise potential user-base. In practical terms this meant compiling for the earliest possible OS version and ensuring that features relying on later OS versions degraded gracefully.

As explained in Section 4.1, users were prevented from starting the game without first stating that they had read and understood the terms and conditions, which explained the nature of the trial and the data that would be logged about their use of the system. However, when speaking to participants, it emerged that none of those interviewed had understood the game was part of an academic trial. A task was subsequently presented to users, further explaining the nature of the trial and asking whether they had understood this, with 70% responding that they had not. This mirrors findings by the Federation Against Software Theft [16] where the percentage of users who reported reading EULA’s on the desktop was 28%, with 72% routinely agreeing to them “without taking any notice of exactly what they are agreeing to.” This highlights a potential ethical issue for all researchers distributing software in this manner as, opposed to a traditional face-to-face handover where participants’ understanding can be gauged and explanations repeated or re-worded as necessary, the understanding of remote participants is assumed on the basis of clicks on a checkbox and can only be verified after they have become involved in the trial.

6 Conclusions and Future Work

We have described running a worldwide trial of Hungry Yoshi, and the means by which this application was distributed to users, on a scale beyond that usually found in ubicomp field trials. A central aim was to push the upper limit on the number of participants as far as we could while still combining quantitative and qualitative approaches in ways that usefully and efficiently fed into the redesign process. We used a distribution method that made the system available to the general public, comprehensive system logging, a means of interacting with users that was integrated with the user experience of Yoshi, and interaction via a social networking web site. The benefits of such mechanisms include a significant reduction in the effort and monetary cost of field trials—particularly the cost of devices for such field trials—as well as an increase in the numeric and geographic scale of the user-base.

The worldwide nature of the trial meant that we had to adapt our tools and methods to maintain awareness of participants. We described how we used quantitative and qualitative assessments to assess the activity and engagement of our user-base, and how we used this to perform targeted interaction with participants, how that interaction took place on a variety of scales, and how we embedded feedback mechanisms within the system and encouraged their use. The Facebook social networking site served as a means to contact users, to give them awareness of other users' activity, and as a means for them to interact with each other. In combination, these features let us run a trial involving a very large number of participants for a long period of time, and yet have relatively quick redesign cycles set within that process. We offer these summarising points for researchers taking a similar approach:

- *Expect low percentages of uptake and participation.* Software on mobile devices has become a ‘disposable’ form of entertainment; expect your software to be treated in the same manner as any other.
- *Be inclusive.* In order to maximise user engagement, lower technical and social barriers to participation not relevant to research issues.
- *Stay in the application.* Communication within the bounds of the application is more acceptable to users, and therefore achieves a much greater response rate. We found an order of magnitude less participation for every step ‘out of the game’ users were asked to take.

In our future work, we will be exploring further ways to enhance users' engagement in reporting problems, proposing new design suggestions and discussing game play and game development. We will offer means for a user to use Facebook to gain access to data collected from his/her device, and consequent analyses and comparisons, and thus create a resource to change his/her own system use and to participate further in the design process. We are also aware of the way that some of our mechanisms may be particular to games, such as tasks that users are motivated to carry out for game advantage. We are therefore considering how to generalise this mechanism to other application areas. Finally, we are exploring breaking up our applications into software components that can be flexibly combined, so as to support finer-grained updates and to support users' customisation of software configurations appropriate to their contexts, preferences and behaviours. We see such customisation as particularly appropriate given the variety of contexts and uses that become open to study as techniques

such as those presented in this paper allow us to increase the scale of system trials beyond the limits of currently standard methods and techniques.

Acknowledgments. We thank Louise Wright for translation services, and the other members of the Social/Ubiqitous/Mobile Group for their collaboration: John Ferguson, Phil Gray, Stuart Reeves and Scott Sherwood. This research was funded by UK EPSRC (EP/F035586/1).

References

1. Rogers, Y., et al.: Why it's worth the hassle: The value of in-situ studies when designing UbiComp. In: Krumm, J., Abowd, G.D., Seneviratne, A., Strang, T. (eds.) UbiComp 2007. LNCS, vol. 4717, pp. 336–353. Springer, Heidelberg (2007)
2. IDC, Worldwide Converged Mobile Device Market Grows 39.0% Year Over Year in Fourth Quarter,
<http://www.idc.com/getdoc.jsp?containerId=prUS22196610>
3. Zhai, S., et al.: Shapewriter on the iPhone: from the laboratory to the real world. In: Proc. ACM CHI Extended Abstracts, pp. 2667–2670 (2009)
4. Bell, M., et al.: Interweaving Mobile Games with Everyday Life. In: Proc. ACM CHI, pp. 417–426 (2006)
5. Hall, M., et al.: Adapting Ubicomp Software and its Evaluation. In: Proc. ACM Engineering Interactive Computing Systems, pp. 143–148 (2009)
6. O'Neill, E., et al.: Instrumenting the city: developing methods for observing and understanding the digital cityscape. In: Dourish, P., Friday, A. (eds.) UbiComp 2006. LNCS, vol. 4206, pp. 315–332. Springer, Heidelberg (2006)
7. Reades, J., et al.: Cellular Census: Explorations in Urban Data Collection. *Pervasive Computing* 6(3), 30–38 (2007)
8. Chin, A.: Finding Cohesive Subgroups and Relevant Members in the Nokia Friend View Mobile Social Network. In: Proc. Social Computing, pp. 278–283 (2009)
9. Licoppe, C., Inada, Y.: Emergent Uses of a Multiplayer Location-aware Mobile Game: the Interactional Consequences of Mediated Encounters. *Mobilities* 1(1), 39–61 (2006)
10. Crabtree, A., et al.: Supporting Ethnographic Studies of Ubiquitous Computing in the Wild. In: Proc. ACM DIS, pp. 60–69 (2006)
11. Froehlich, J., et al.: Voting With Your Feet: An Investigative Study of the Relationship Between Place Visit Behavior and Preference. In: Dourish, P., Friday, A. (eds.) UbiComp 2006. LNCS, vol. 4206, pp. 333–350. Springer, Heidelberg (2006)
12. Carter, S., et al.: Support for Situated Ubicomp Experimentation. In: Proc. ACM CHI, pp. 125–134 (2007)
13. Morrison, A., et al.: Using Location and Motion Data to Filter System Logs. In: LaMarca, A., Langheinrich, M., Truong, K.N. (eds.) Pervasive 2007. LNCS, vol. 4480, pp. 109–126. Springer, Heidelberg (2007)
14. Common European Framework of Reference for Languages,
http://www.coe.int/T/DG4/Linguistic/CADRE_EN.asp
15. Facebook Statistics,
<http://www.facebook.com/press/info.php?statistics>
16. FAST, Federation Asks: Do you know what you're agreeing to?,
<http://www.fastiis.org/resources/press/id/304/>

Studying the Use and Utility of an Indoor Location Tracking System for Non-experts

Shwetak N. Patel^{1,2}, Julie A. Kientz^{3,4}, and Sidhant Gupta¹

¹ Computer Science & Engineering, ² Electrical Engineering, ³ The Information School,

⁴ Human Centered Design & Engineering

UbiComp Lab, DUB Group, University of Washington

Seattle, Washington, 98195

{shwetak@cs., jkientz@, sidhant@cs.}washington.edu

Abstract. Indoor location tracking systems have been a major focus of ubiquitous computing research, and they have much promise to help in collecting objective, real time data for applications and supporting studies. However, due to their typically difficult and time consuming installation process, few have explored the extent to which they can be used by non-experts. In this research, we studied how one location tracking system, PowerLine Positioning, could be used by non-technology expert rehabilitation researchers to study the mobility patterns of wheelchair users in their homes. We determined that indoor location tracking systems are not only usable by non-experts, but they can also be useful in allowing them to achieve their own research goals of obtaining objective mobility data. Based on the results, we provide areas for future exploration and implications for designers of location-based and other types of sensing systems which aim to be end-user deployable.

Keywords: Location, indoor location sensing, end-user deployable, accessibility, wheelchair users, PowerLine Positioning, user study.

1 Introduction and Motivation

The ability to sense a person's location, both indoors and outside, has been a long-term goal of technology designers in pervasive and ubiquitous computing. The promise of knowing precisely where a person may be at any given time has many potential uses, from context-aware computing, to location-based services [19], to tracking and monitoring behavior [3, 4, 24]. A number of researchers, both within the ubicomp community and beyond, have an interest in using location tracking systems as a means of collecting data they can use for assessing people's behaviors, activities, and whereabouts. This may be useful for application designers who wish to provide location-based services, but also for researchers hoping to better understand people's behaviors using a more objective means.

One difficulty with current sensing systems, especially indoor location sensing, is that the systems typically require extensive setup and have significant installation burdens on end-users [4]. Thus, their use often requires an expert to help install the application or requires significant training on the part of the end-user. These aspects

limit the number of locations that a sensing system may be deployed as well as the number of people that can be studied. Because of these factors, the usefulness and utility of location sensing systems by non-technical users has been underexplored.

In this research, we wanted to determine whether the promise of easy-to-deploy indoor location systems could live up to the usefulness predicted by many location system designers. Thus, we explored whether one indoor location tracking system, designed to be easy-to-deploy and low-cost, could be used by a set of non-expert users to achieve their goals for location tracking research. In particular, we studied a group of rehabilitation researchers who used the PowerLine Positioning (PLP) [25] system to track the mobility patterns of wheelchair users within their homes to augment interviews having the goal of uncovering accessibility barriers and everyday life experiences. In particular, we focused on how well end-users were able to install, calibrate, and use PLP in the homes of the wheelchair users they wanted to study. We also studied the usefulness of PLP in achieving the goals of the researchers by identifying whether researchers were able to uncover more information about accessibility barriers by using PLP in combination with interviews than interviews alone. Overall, we found that PLP was easy to deploy and maintain by the rehabilitation researchers and helped them to uncover more barriers to accessibility in the wheelchair users' homes than interviews alone.

The research we present in this paper helps motivate one particular need for indoor location tracking systems as well as show how a ubiquitous sensing system can be designed to support use by the non-expert end-user. We provide a number of implications that other technology designers can use to design their systems to be both usable and useful by end-users. These implications can also be used for the design of sensing systems beyond just location tracking. This helps bring the field of ubiquitous computing closer to achieving the goal of sensing on a larger scale by creating guidelines for easy-to-deploy and maintain sensing systems.

The rest of this paper is organized as follows. We begin with a discussion of the related work, where we describe what motivated this work and where this work fits within the larger scheme of the pervasive computing literature. We next describe the design of our study, including details about the PowerLine Positioning technology we had users deploy in this study as well as the research questions the rehabilitation researchers wanted to address. Next, we describe the details of the study findings, including data on the ease of use and usefulness of the technology for the rehabilitation researchers. We follow the findings with a discussion of the results and describe implications and lessons learned from this process and then end with the conclusion.

2 Related Work

In this section, we outline the related work for this research and how it builds upon existing technology and studies. In particular, we discuss indoor location tracking systems and their challenges for domestic use, studying human activity in the home, and studying end-user deployment of sensing systems.

2.1 Indoor Location Tracking Systems and Their Challenges

Indoor positioning has been a very active area of research in pervasive and ubiquitous computing for the past decade, and many commercial systems are beginning to emerge. Several characteristics distinguish different solutions, such as the underlying signaling technology, line-of-sight requirements, accuracy, and cost of scaling the solution over space and over number of items [31]. The first indoor solutions introduced new infrastructure to support localization [1, 14, 22, 27]. Despite some success, as indicated by commercialized products [10, 12, 16, 30, 32], the cost and effort of installation are major drawbacks to wide-scale deployment, particularly in domestic settings. Thus, new projects in location-based systems research reuse existing infrastructure to ease the burden of deployment and lower the cost. The earliest demonstrations leveraged 802.11 access points [6, 8], and more recent examples explore Bluetooth [20] and wireless telephony infrastructure, such as GSM [19, 23] or FM transmission towers [18]. A concern is that individuals may not be able to control the characteristics of this infrastructure and the operational parameters of the infrastructure may change without warning, resulting in the need to recalibrate. The desire to control the infrastructure and to scale inexpensively inspired the work on the Power-Line Positioning system [25], which we used in this research.

Deployment time and ease-of-use are other essential considerations for indoor location systems, especially for studies in domestic settings. Investigators have limited time they can spend in a participant's home, thus the entire installation process must be as short as possible. In addition, technical expertise can also vary greatly, so an easy-to-use solution is always desirable. One way to address this challenge is to minimize the number of components used in the system, which is the case of the PLP system. Studies have also shown that homeowners are concerned with the appearance of their home after adding any additional instrumentation [7, 15]. PLP's minimal components made it an ideal system for us to use in this study.

2.2 Studying Human Activity in the Home

With the advent of new, affordable technologies, there has been a trend in research to shift from building technology to supporting office and home life. Abowd and Mynatt point out a need for studying domestic settings to inform the design of new technologies [1]. Edwards and Grinter echo similar sentiments in that people are using technologies in new and interesting ways in the home [11]. Thus, a key research problem for designing for the home is first to study the home's everyday workings, such as how people live in the home, what they do, and the role that technologies play.

The initial foray in studying the home has been with ethnography. For example, Crabtree *et al.* present a series of ethnographic studies that aimed to uncover communication routines and how people use particular spaces in the home [9]. They provide guidelines for placing technology in appropriate locations in the home. More recent work has looked at collecting empirical evidence for studying the domestic space. For example, Intille *et al.* presents techniques for acquiring data about people, their behavior, and their use of technology in a natural setting [17].

With the proliferation of portable electronic devices in the home, researchers are interested in studying the complex interactions between household residents and their devices. Aipperspach *et al.* [3, 4] looked at using sensor-based visual records of the physical movement of people and devices to facilitate in-depth discussion during interviews, but they also report challenges in installing the Ubisense location tracking system, which impacted the number deployments. Rowan and Mynatt installed strain sensors on the underside of the first floor of an elder's home to deploy their Digital Family Portrait application [28]. By detecting the weight of a person standing on the floor, these sensors allow the Digital Family Portrait to portray movement information in the home.

Tapia, *et al.* describe MITes (MIT environmental sensors), which are low-cost, wireless devices for collecting real-time data of human activities in natural settings [29]. The system includes five wearable sensors: on body acceleration, heart rate, ultra-violet radiation exposure, RFID reader wristband, and location beacons. Patel *et al.* demonstrated the use of Bluetooth for tracking the proximity of users to their mobile phone to study their affinity to their mobile phone and reasons for separation [24]. Philipose, *et al.* present the use of an RFID-enabled glove to monitor activities of daily living [26]. A person wearing the glove interacts with RFID-tagged objects, and the system recognizes activities based on interactions with objects. Thus, many of the current sensing approaches aim to address particular behaviors and location is often implicitly inferred. Just sensing alone cannot always gather meaningful information on people's activities, and thus must be coupled with an annotation or survey procedure. This line of research shows a need for developing and testing robust, scalable sensing systems for the home, as we have done with this work. In addition, we highlight the value of augmenting self-report with real-time location sensing data.

2.3 Studying End-User Sensor Deployments

Researchers have explored the acceptance of sensors in the home as well as end-user deployment considerations, which provided us with some guidelines for designing our location tracking solution. Hirsch, *et al.* examined the social and psychological factors that influence the design of elder care sensing systems and applications [15]. Among their findings is a concern that technology may be rejected if it detracts from the aesthetics of the home. Beckmann, *et al.* presented a study of end-user sensor installation and reaction to sensors in the home [7]. They had end-users install vibration sensors, in-line electricity monitoring sensors, motion detectors, cameras, and microphones. They found that end-users made a variety of errors, often due to the directional requirements of sensors or uncertainty over exactly where a sensor needs to be positioned. They also found many negative reactions to the intrusion of sensors into the living space, including objections to the potential for damage caused by the adhesive used for installation, concerns that sensors were placed in locations accessible by children or pets, and objections to the placement of cameras and microphones in the home. We use some of these principles in the design of our deployable system in addition to offering new insights in building non-expert or end-user deployable location tracking solutions for home.

3 Study Design

In this section, we describe the overall design of our study. We begin with the motivation and design of the rehabilitation researchers' study and the questions they aimed to answer, describe our method for studying them, and then discuss the technology we used to support the rehabilitation researchers. The overall aim of this our study was to answer the following research questions:

- Can an indoor location tracking system be easy enough to deploy for non-experts, and what is the typical deployment time for a standard home?
- Can an indoor location tracking system provide objective, empirical data for location-based studies in the home?
- Does using automatically-sensed location and mobility data facilitate a richer interview process that produces higher quality data?

3.1 Motivation for Studying Mobility of Wheelchair Users in the Home

Increased activity and participation for people with disabilities is a goal of the U.S. Americans with Disabilities Act (ADA) [5] and the New Freedom Initiative [21]. The aim is to reduce environmental barriers and increase access to assistive technologies in order to increase the ability of people with disabilities to integrate into the community, have a sense of autonomy, and lower dependence on societal resources. The assumption for this population is that wheelchairs are necessary for mobility, and mobility is the means to performing activities and community participation. However, in order to dress, eat, or bathe in a wheelchair, the home environment needs to be accessible (*e.g.*, wide enough doorways, wheelable ground surfaces, *etc.*). Studying and understanding the mobility patterns of wheelchair users in their homes can provide useful insights on where environmental barriers exist, how better to design assistive devices, and how to improve the architecture of homes and offices. However, collecting this data is a difficult task.

New technological methods are needed to understand activity and participation in the everyday lives of wheelchairs users, especially in the home. The development of objective indoor measures is critical to understanding how people use mobility devices in the home and can be used to document where, when, and how people are using these devices. In addition, it can help understand how specific environments in the home can facilitate or hinder a person's use of a particular device or the performance of a specific activity. In the rehabilitation research community, current measurement of indoor activities among wheelchair users in the home has been limited to self-report questionnaires, such as the Home Accessibility Survey [13]. Disability researchers have also used diaries to gather mobility problems when they occur. However, researchers have found that many incidents are missed with both of these methods. Indoor tracking technology can provide simple, automatic, and objective data about the activity and participation of individuals in a space.

Table 1 shows some the types of questions that researchers are interested in gathering about wheelchair users. The questions are based on current literature in the area. For each question, we briefly highlight how automatic location tracking can play an important role alone and when used in combination with interviews.

Table 1. Questions the rehabilitation researchers are interested in and how location tracking technology can help answer them

1. Where do people go to perform what activity?
<i>Location data can show where people tend to spend a lot of time, and the interview can probe the participants about what activity is going on at those times.</i>
2. How do people who use mobility devices make use of spaces in the home, and how does it differ from non-disabled family members?
<i>Location data can show traces of where disabled and non-disabled members tend to go.</i>
3. What mobility devices do people use for a particular activity (e.g., walker to enter the bathroom, shower chair in the bath, and wheelchair in the hallway)?
<i>Tagging all mobility aids will give information about which are used in which parts of the home. Participants can be queried about particular situations to determine the reason for the transition.</i>
4. What is the frequency and duration of mobility device (e.g., walker, cane, wheelchair, etc.) use in each room?
<i>Location data can provide this information automatically and more accurately than self-report.</i>
5. What routes do individuals take throughout the home? How have people adapted their homes (or not) and how does that impact mobility device use?
<i>Location data can be aggregated to show time varying route information for each individual. The interview process can potentially reveal why certain routes are taken, such as the result of an environmental barrier.</i>
6. What parts of the home are completely inaccessible?
<i>Location data can show the parts of the house where people rarely go, and the interview can determine the reasons (e.g., inaccessible or just not used).</i>
7. What are key facilitators in people's homes (e.g., caregivers, devices, furniture, etc.)?
<i>Location data can show the routes people take and if an aid was used. If it is not shown being used, then during the interview the participant can be probed about other types of mobility assistance they may use (e.g., help from a family member).</i>

In large clinical trials, it is often necessary to recruit participants in distant locations. In addition, because of the individual's mobility disability, they may have limited ability to deploy any technology themselves. A deployable system has to be comprised of minimal components that can be installed by anyone, such as by a caregiver or family member. Presumably, researchers would conduct many simultaneous studies to produce a rich and generalizable result, which argues for a cost-effective and easy-to-deploy solution. PowerLine Positioning is an appealing choice to address this need, where the infrastructure requirements are two plug-in modules, and the calibration step consists of a house walkthrough.

3.2 Technology Used in Study

The indoor location tracking system we chose for this study is the PowerLine Positioning system we previously developed [25]. PowerLine Positioning is an affordable, whole-house indoor localization system that works in the vast majority of households, scales cost-effectively to support the tracking of multiple objects simultaneously, and does not require the installation of any new infrastructure. The solution requires the installation of two small, plug-in modules at the extreme ends of the home (e.g., the upstairs northwest corner and the basement southeast corner). These modules inject a mid-frequency signal throughout the electrical system of the home. Simple receivers, or positioning tags, listen for these signals radiated off the power line and wirelessly transmit their positioning readings back to the environment.

We re-engineered and built a deployable version of the PLP system, which conforms to the U.S. FCC Part 15 regulations. In addition, we incorporated a multi-staged tuned tag design (see Figure 1) and used different power line transmission frequencies: 500 kHz and 600 kHz (see Figure 2). This new design yielded better performance and resolution (86% classification at 1-meter regions) in our test deployments.

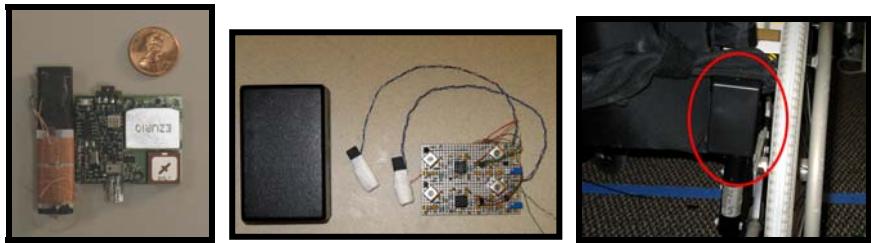


Fig. 1. Left: Redesigned and deployable PowerLine Positioning tags. **Middle:** The encasement used for larger devices, such as wheelchair and walkers. The larger case housed a higher capacity battery. **Right:** The tag installed on a user's wheelchair.

The new location tags we developed featured a zigbee wireless backchannel that reported a 16-bit unique ID, two 12-bit signal values, and a single bit indicating if the button on the tag is pressed back to the basestation in the home. The RF receiver connected to the personal computer was able to receive data from up to 25 tags (base station limitation). An application running on the laptop parsed the data, handled the fingerprinting algorithm, and provided location services to the visualizer. The tag had an on/off switch and a single position push button. The button was used to indicate a special action to the remote computer, which was used to indicate that the tag is in site survey or calibration mode. The tags also incorporate motion detection, so the tag will go into a sleep mode if no motion is present for 30 seconds and reactivate itself on the next motion event. This approach greatly reduced the overall power consumption of the tag. With the tag duty cycling 40% of the time, the tag could easily last the entire duration of the study using a 750 mAh lithium ion battery source.

To convey the location data for the interviews, we developed a simple visualization tool that used the data provided from PowerLine Positioning. The visualization allowed researchers to enter a timeframe and view either the concentration of activity levels in a given area in the house for a particular mobility aid or participant or view the routes that users or mobility aids traveled throughout the home (see Figure 3). The routes could also be animated to play back the exact path. The visualization tool

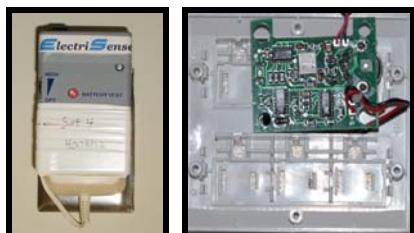


Fig. 2. Left: The signal generating plug-in modules. **Right:** Inside back cover of the outlet housing the signal generating circuitry.

provided a simple means to scan a large amount of location data in a meaningful way, by superimposing graphics on top of a floor map of the participant's home, drawn by rehabilitation researchers using a free online mapping tool.

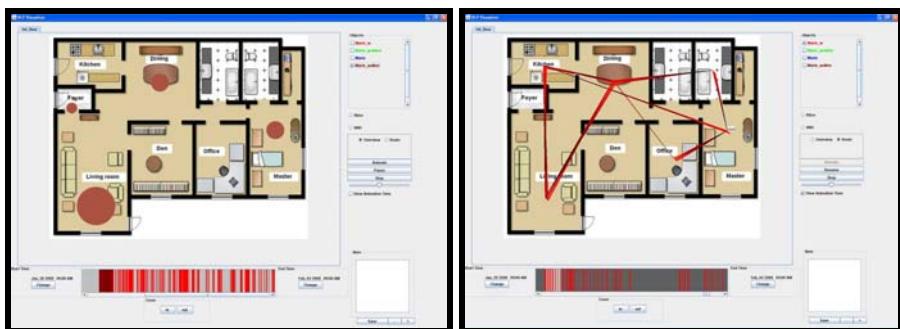


Fig. 3. Visualization of PowerLine Positioning data used during the interviews. **Left:** Size of red dot provides length of time tracked entities were at particular locations. Colored vertical bars represent movement of the corresponding entity. **Right:** Mobility traces or routes of the tracked objects and people. Black bounding bars on the timeline indicate how trail length shown on map. The routes are drawn as a line segment to show the origin and destination.

3.3 Study Details

The deployment study involved rehabilitation researchers studying the mobility patterns of four different households (see Table 2). The participants were recruited by the rehabilitation researchers and were selected through their patient pool of wheelchair users by sending out recruitment emails and letters. Each household was enrolled in the study for 6 weeks during which 7 interviews (3 current practice self-report and 4 prompted-recall) were administered on an approximately 1-week basis (see Table 3). Rehabilitation researchers also conducted interviews regarding the obtrusiveness and acceptance of the technology on the days of instrumentation installation and removal. The installation of PLP was carried out entirely by two rehabilitation researchers while we played only an observer role. The backgrounds of the researchers were in anthropology and design, and neither had previous experience with setting up or using location tracking systems. For each of the four deployments, a rehabilitation researcher installed PLP, and we also evaluated the installation and maintenance for each. We trained the two researchers prior to deployment, which involved a 30-minute tutorial and an installation example in a laboratory. During this tutorial, they were asked to install the system themselves under our supervision (setup and install the hardware for tracking 4 tags and calibrate the software). A PLP installation manual was also provided to the installers to read beforehand if they chose to do so. For the actual deployment, each of the two installers installed PLP in two homes. We timed how long each installation took and interviewed those conducting the installation to determine ease-of-use and problem spots during the installation.

For the study, we attached a location tag to each mobility device used and gave a tag to each member of the household. We built custom mounts that allow easy

attachment to round surfaces, such as the frame of a wheelchair or a walker (see Figure 1). Individuals were asked to wear their tag around the neck on a lanyard. The batteries for each tracking tag were replaced or recharged during each weekly interview. However, we found that weekly recharging was unnecessary, because the tags ended up lasting over one month.

For the current practice self-report interviews, the researchers administered a Home Accessibility Survey [13], which is an interview process that captures the subject's knowledge, comfort, and satisfaction with their mobility aids and perceived environmental barriers to their mobility device usage during the past week. This tool is a hybrid survey the rehabilitation researchers created that synthesizes various well-known interview questions from the disability mobility community. Current research measures on environmental barriers in the home are limited to self-report and how barriers are subjectively experienced by the user. We maintained this process so that we could compare the self-report data with data gathered from the prompted semi-structured interviews based on the mobility pattern data provided by the tracking system. For the prompted-recall interviews, rehabilitation researchers reviewed the position traces using the visualization tool (captured with PowerLine Positioning, see Figure 3) with the participant from the previous day and "prompted" them with questions based on the mobility data. Because of the richer data, these interviews were typically limited to reviewing the participant's prior day. The interviews were scheduled such that different days of the week were gathered (*e.g.*, weekend vs. weekday). Table 3 shows the duration of the study and the interview schedule.

The aim was to compare the level of detail and quality of data researchers could obtain by using the sensed data as part of the interview process, in contrast to relying on self-report alone. For example, one metric of success was the determination of the number of environmental barriers to mobility for that person. Thus, we compared how many more barriers were found with the PLP-based interviews than self-report data to assess its effectiveness.

Table 2. Demographic information for each household of the wheelchair mobility participants

Household ID	Gender	Age	Profession	Mobility Aids Used	Years Using Chair	Home Size (ft ² /m ²)	Number of Rooms in Home
1	M	38	IT Specialist	Powered wheelchair	30	1100/102	5
2	F	62	Consultant	Powered wheelchair, Walker, Grabber	26	1600/149	8
3	M	51	Public service business owner	Manual wheelchair, Manual sports wheelchair	28	1800/167	8
	F	48	Sales associate	N/A	N/A		
4	F	33	Product manager	Manual wheelchair, Walker	12	2400/223	10
	M	36	Engineer	N/A	N/A		

Other quantitative measures of activity gathered through the self-report questionnaire included metrics such as the length of time they spend out of the bed, the frequency that they move from one end of the home to the other, the percentage of the time participants spend using each mobility aid, and where they use the mobility aids. With the location data, we aimed to evaluate the accuracy of self-report responses using these objective measures. Finally, we interviewed the investigators conducting these studies to evaluate the ease-of-use of the pattern traces and their usefulness during the interview process.

A total of 6 different interviewers from the research team conducted the interviews with the four mobility participants (or households). For a given mobility participant, the interviewer that conducted the HAS-based interviews (current best practice) was different from the interviewer conducting the PLP-based interviews using the mobility traces. The reason for this was to ensure that the two interview processes did not bias each other. In order to address the issue of the differences between the two interviewers, we attempted to recruit interviewers that had similar experience levels both in conducting interviews and in the disability research community. In addition, we alternated the roles of the interviewer for each participant to counterbalance the interviews and varied when the non prompted-recall interview was administered during the 6-week period.

Table 3. Timeline of interviews conducted with members of each home during the course of the 6-week study. H = HAS-based interview, P = Prompted-recall, A = Interview on the acceptance and obtrusiveness of technology.

	Week of Study					
	1	2	3	4	5	6
H1	H,A	P	H,P	P	H	P,A
H2	P,A	H,P	P	H	P	H,A
H3	H,A	P	H,P	P	H	P,A
H4	P,A	H	P	H,P	P	H,A

One important consideration was the potential concern participants may have regarding privacy, especially in the home where it is a very personal space. Although the location data did not produce the same level of detail as video recordings, it was still important to be sensitive to what the participants were willing to reveal. We addressed this concern in two ways. The first was by giving the participants the ability to stop collecting location data at any moment by pressing the button on their location tag. Pressing the button again would restart data collection. The second was by creating a trusting relationship with the interviewer during the data review and interview process. This was accomplished by initially sharing with them their mobility trace data. Trust was also established by making participants a partner in the research process and having them drive the interview by asking them to walk through their day with the interviewer. The interviewer in turn asked more specific questions based on what the participant chose to reveal and what they saw from the mobility trace. In addition, the interviewers were instructed to be sensitive to questions and/or issues participants might find invasive and uncomfortable.

4 Results

In this section, we describe the findings from the study we conducted. We first describe the usability of the PowerLine Positioning system by the end-users. We then outline the results of the study on the usefulness of the PLP system in achieving the goals of the researchers, as compared to traditional data gathering means.

4.1 Usability of the PLP Tracking System by Non-experts

We assessed both the performance of the PowerLine Positioning system in these deployments as well as the ease of deployment by observing the installation procedure and interviewing the installers. Overall, both systems were successfully deployed in all four homes. We merely observed the installation procedure and did not directly assist them in the installation in any way. When the tracking system reported at least 95% accuracy from 20 random locations throughout the home, the installation process was concluded. PLP took an average about 32 minutes to install ($H_1 = 25$, $H_2 = 29$, $H_3 = 41$, $H_4 = 33$), which is encouraging considering the overhead other approaches would have had if there was additional hardware that needed to be installed in the home. This time includes the planning, physical installation, calibration, and testing of the system. Most of the time was attributed to the site survey or calibration. Homes 1 and 2 were installed by one person and Homes 3 and 4 by a second.

The installers appreciated the minimal amount of devices that needed to be deployed in the home since PLP required very little hardware that actually needed to be physically installed. During the study, the installers conducted performance tests when meeting with the participants for their interviews to determine whether recalibration was necessary. A recalibration was determined to be necessary if more than 10% of the tests failed to produce a correct position reading. Only one installer reported having to conduct another site survey (Home 1) in the middle of the study. The reason was during the interview process, she noticed two regions of the home were not being tracked, thus she needed to update the signal map with a denser survey. Home 1 required part of the living room and master bedroom to be resurveyed during the second week. The other 3 homes required no additional surveys.

To assess the researcher's proper maintenance and overall installation of PLP, we evaluated the overall accuracy of the system for the entire duration of the study. For this, we had the wheelchair users manually provide labeled ground truth data throughout the day by simply pressing a button on a wireless module placed at fixed location in the house (10 per home). We typically put these near frequented areas, such as the dining room table, office or computer desk, night stand, and coffee table. We asked them to simply press the button when they noticed it and had the opportunity. Since we knew the exact position of those buttons, we were able assess the performance of PLP (the classification accuracy for that sub-room) at those known locations. PLP correctly indicated the person's location (within a 2 meter circle) at the time the button was pressed. Over 97% (507 out of 523) of the button presses were identified correctly with PLP across all homes.

4.2 Utility of Indoor Location Tracking System

The four households that were enrolled in the study each had researchers conduct four PLP-based prompted-recall interviews and three prior practice self-report interviews. The prompted-recall interview consisted of a 1-hour meeting with each participant and a walkthrough of his or her previous day (two days if time allowed) using the PLP tracking data. The participants were allowed to dictate what was shown on the tracking interface to talk about any detail they chose, but the interviewers were instructed to follow the interview guide as much as possible. The tracking data was used to help prompt the participants about interesting situations that might have occurred with their mobility aid. In addition, the data was also used to encourage the participants to reflect on their usage of various mobility aids. The non-prompted interview was conducted using an adapted version of the rehabilitation researcher's current practice surveys (the HAS). These interviews also lasted about one hour, and the interviewers were asked to follow the interview guide. The interview was very similar to the prompted-recall interviews except that the tracking data was not available. The interviewers asked each participant to reflect on their previous two days during their interview, although they were not limited to that.

Each interview was audio-recorded, and the PLP tracking software logged when various features of the software were used. The interviewers also took notes during the interview. After the completion of the study, the interviews were transcribed for further analysis. We also analyzed the PLP tracking data to extract quantitative measures, such as time spent in each room, percentage of time spent in each room throughout the day, *etc.*, to compare against the participants' recall of that information.

The interview notes and transcripts were used to extract relevant statements and discussion points generated during the interviews, which in turn were used to produce themes that emerged from all the interviews relating to mobility problems. Two rehabilitation researchers independently categorized the statements in the transcripts and notes to determine the themes. The two coders produced a total of 19 themes, eight of which were common across the two coders. Thus, 11 unique themes were included after discussion and resolving overlaps between the different themes. A third independent coder re-categorized the statements using these 11 themes and we calculated inter-rater reliability using the categorizations from the three coders using two measures: observed agreement and Cohen's Kappa (see Table 4).

Table 4. Inter-rater reliability for each theme: (1) Observed agreement, measured by agreements divided by total number of statements coded and (2) Cohen's Kappa. Measures are between 0 and 1, with 1 indicating perfect agreement between coders.

Cluster	1	2	3	4	5	6	7	8	9	10	11
Observed Agreement	.96	1	.96	.98	.95	.95	1	.96	1	1	.95
Cohen's Kappa (k)	.92	.96	.83	.95	.79	.80	.96	.79	.94	.96	.80

The following themes emerged after the data analysis:

1. *Mechanical problems*: physical problems with the mobility aid itself, such as a broken wheel, faulty brake, etc.
2. *Mobility aid form factor or design problems*: the aid does not serve its purpose or intended function
3. *Doorway, hallway, or threshold barriers*: problem in locomotion in the home because of environmental barriers
4. *Reach problems*: items of interest being out of reach
5. *Level access problems*: includes accessing items that are hard to maneuver to, which can result from not being able to rotate the wheelchair, cluttered room, etc.
6. *Exercising*: tasks relating to regaining mobility strength, such as home physical therapy
7. *Safety concerns*: afraid of falling or not being confident enough to go to a particular region of the house or perform a particular task
8. *Person assistance*: task requires assistance from an able-bodied individual
9. *Floor conditions*: the characteristics of the floor contribute to mobility concerns, such as using a walker on carpet or slippery floors
10. *Self-conscious*: reluctant to show they used a mobility aid
11. *Medical procedures*: recent medical procedures or changes in health affecting overall mobility

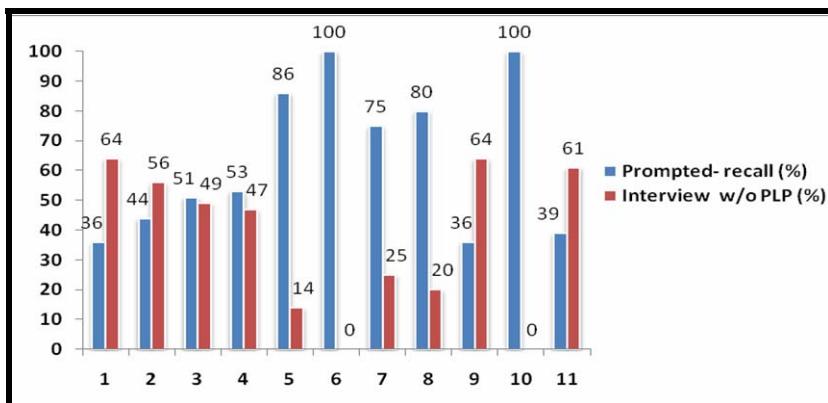


Fig. 4. Percentage of discussion points resulting from prompted-recall and non-promoted-recall interviews for each theme

Both interview methods (prompted-recall and non-prompted-recall) produced responses in 9 of the 11 themes (see Figure 4). However, two themes (Theme 6 and 10) only emerged from the prompted-recall data. In addition, a higher percentage of discussion points relating to Themes 5, 7, and 8 were produced from the prompted-recall data. Thus, there were some clear advantages to having the tracking data available during the interview process. For example, in the case of exercising, participants often talked about using a particular mobility aid for the purposes of strengthening their legs or muscles. However, it was not until participants actually saw their activity data did

they recall this detail. Similarly, seeing their tracking data prompted discussions about other individuals having to help with a particular task. Participants also discussed situations where they did not take a particular route in their home or use a particular aid in certain parts of the home because they were afraid of falling (Theme 7). Another interesting result from the prompted-recall data was that participants not only talked about physical barriers in their environment, but also social pressures (Theme 10). For example, there was one instance where the data showed that a participant started to use a different aid that was not normally used. When she saw this, she stated that she did not want to her grandchildren to see her in a wheelchair, so she made a conscious effort to use the walker during their visit. Another participant reported using a manual wheelchair when his friend would come over, who also used a manual wheelchair.

In addition to counting the number of themes that emerged with each approach, a second coding scheme was introduced to rate the quality of the coded discussion points as determined by the needs of the rehabilitation researchers. Two coders rated each of the 113 statements or discussion points around that statement with a rating of 1 or 2. A value of 1 referred to a statement that was mentioned, but the participant did not engage in supporting details or examples during that discussion, while a value of 2 was given to a discussion point that involved the participant giving specific details. We also calculated a percentage agreement and Cohen's Kappa for the rating scheme for the 113 statements, which resulted in an observed agreement of .96 and a Cohen's Kappa of .88. The aim of this coding scheme was to determine the number of rich discussion points that resulted from using the prompted-recall method compared to the standard interview. In general, we saw a higher quality level of statements gathered using the prompted-recall ($\sigma = 1.85$) than we did with the interviews alone ($\sigma = 1.4$), which a two-tailed T-Test showed to be significant ($p < 0.01$). The higher rating of the prompted-recall interviews could be a result of the participants having something to explain or narrate when using the tracking data. In the self-report data, it was often the case that participants rarely remembered details around their actions during the prior days. One participant referred to the tracking data as, "the next best thing to a video camera without a camera," alluding to the usefulness of the context it offered during the interviews.

5 Discussion

The results of this study show that the use of an indoor location tracking system is feasible and acceptable for study participants and helps to meet the needs of non-experts. The study enabled us to uncover interesting results and implications for designers of other sensing systems, as well as those conducting studies of sensing systems in the home. Here we discuss the value of location-based sensing and the future avenues that can be explored, the implications we found for conducting these types of studies, and describe the limitations to this study.

5.1 Value of Sensing for Non-experts

One of the main contributions of this study was to show that location-based systems can be both usable and useful to non-experts. We believe this opens a number of doors for researchers and helps to validate much of the location-based technology research. Beyond the domain of sensing the location of wheelchair users, we believe

there are other application areas that are made possible by this line of work. For example, research applications in eldercare [17] have made use of location data as a peace-of-mind application. These applications could be simplified and are feasible for end-users to deploy in their homes with the PowerLine Positioning system. Another potential application area is for architects who wish to study the use of built environments. Occupants could wear tags to allow designers to see which parts of homes are used and which are not to help determine how spaces could be redesigned or renovated to better use the space. Finally, this type of technology could benefit families of children with special needs, who could consider using indoor tracking systems to identify which rooms children are in when they exhibit different behaviors, which could help determine the causes of the behaviors and better address their needs.

5.2 Implications and Considerations

As a result of this work, we determined a number of implications for designers of future sensing systems for the installation and use by non-expert users and researchers who may want to use these techniques to study human behavior.

- *Visualization of Data* – One of the most important aspects of a sensing system designed for non-experts is an effective means of visualizing the data in a way that is relevant and easy to understand. The tools we built for visualization were rudimentary, and while they allowed the rehabilitation researchers to do the job, more design consideration could make the tools easier to use and understand. Understandable visualization of location data remains an area for exploration.
- *Initial Calibration* – During the site survey, the installers did not get feedback about the performance of their system until the very end of the calibration procedure. Providing feedback about the accuracy should be shown as they did the survey so they know if they need to do a denser survey to achieve the desired level of accuracy. This could help reduce installation time even further.
- *Interview Timing* – One thing rehabilitation researchers noticed about interviews with the prompted-recall data was that participants often became fascinated with viewing their own location traces, especially the first time they saw it. This may add to the length of the interview time during the first visit, since the participant may want to explore their own data. We do not think this is a negative aspect, but should be considered when scheduling prompted-recall interviews.
- *Re-calibration* – In the deployment of PLP for the non-experts, we instructed the installers to place a location tag in a fixed location that would serve as a means for measuring when the system would need to be calibrated. This manual checking worked in our study, but would be helpful if it sent a notification to the installers that a recalibration was necessary. This is especially important for longer-term installations.
- *Data Transparency to the Participants and Privacy* – Participants in the wheelchair mobility study appreciated that they were able to view their own data and see what the researchers could see about their movement patterns. This helped them feel more comfortable and reduce some of the concerns over privacy. We also provided a means for the participants to delete data after it was recorded if they chose, which gave them control over what was recorded. Although this feature was infrequently used, participants expressed comfort in its existence and that the PLP data was much less invasive than cameras.

- *Simplified Data Views* – At some times during the prompted interviews, the data became too overwhelming and confusing for the study participants. Providing ways for researchers to hide some of the complexity at times to make it more understandable to the participants may help make some of the interviews go more smoothly. However, viewing the full data set should always be an option to preserve the data transparency guideline provided above.
- *The Importance of Unobtrusiveness in Domestic Spaces* – Echoing some of the previous work in this space, the participants in the wheelchair mobility study expressed an appreciation for how minimally intrusive the PowerLine Positioning system was to the aesthetics of their home. Most participants agreed that this type of system could remain in their home indefinitely due to its unobtrusiveness.

5.3 Study and Technology Limitations

Although the results of this study were promising in showing the value of sensing and lead to a number of design implications, there were several limitations that we would like to discuss. First, the sample size studied was fairly small, and we were limited to only one group of rehabilitation researchers, who conducted pilot deployments in four households. Thus, the issue of scaling deployments to a large size is still an opportunity to explore. We believe that scaling this study to a larger set of users will be possible, since we found that non-experts were able to install the location-tracking system with minimal training and in about 30 minutes. They were also able to maintain it with very little external technical assistance. Although the non-experts only required minimal training, we would still like to get to the point where the sensing technology could be deployed without professional or expert help, such as by creating a comprehensive installation guide and demonstration video.

With regard to the technology, there were a few problems we encountered that could be improved. For example, positioning tags were still large and did not attach well to the smaller mobility aids, such as canes and grabbers. Thus, some data collected by the system may not be entirely accurate if location tags slip off. In addition, we believe that because PLP is nearly invisible, the non-experts had some difficulty establishing a good mental model of it. In the case of House 1, if they did not feel they were getting a good signal, they would move the plug-in modules to other plugs. Thus, we could provide a better sense of how the system works to end-users. Overall, compliance was still very high across all four participants and they all indicated no major challenges with the tags attached to their primary mobility aids.

6 Conclusion

In this paper, we discussed the usability and utility of a low-cost indoor location tracking system in the context of being used by non-technology expert rehabilitation researchers to study the mobility patterns of wheelchair users in their homes. We have encouraging results that such indoor location tracking systems are not only usable by non-experts, but they can also be useful in allowing non-experts to achieve their own research goals of finding more barriers to access and achieving higher quality, more detailed responses from participants. We hope to entice more work in applying ubiquitous computing technology and building tools to help researchers in other communities wanting to collect objective data about human activity and behavior. In addition,

focusing on the non-expert deployability of these sensing technologies is going to be critical for attaining the scale for ubiquitous and pervasive computing we hope to achieve in the future.

Acknowledgements

We thank the Center for Assistive Technology and Environmental Access (CATEA) for their participation and assistance in this study. We also thank Mayank Goel for his help with developing the visualization tool.

References

1. Abowd, G.D., Mynatt, E.D.: Charting Past, Present, and Future Research in Ubiquitous Computing. *ACM Transactions on Computer-Human Interaction* 7(1), 29–58 (2002)
2. Active Bat. The BAT Ultrasonic Location System (2010), <http://www.c1.cam.ac.uk/research/dtg/attarchive/bat/>
3. Aipperspach, R., Rattenbury, T., Woodruff, A.: A Quantitative Method for Revealing and Comparing. In: Dourish, P., Friday, A. (eds.) *UbiComp 2006*. LNCS, vol. 4206, pp. 1–18. Springer, Heidelberg (2006)
4. Aipperspach, R.J., Woodruff, A., Anderson, K., Hooker, B.: Maps of Our Lives: Sensing People and Objects Together in the Home. EECS Department, University of California, Berkeley. TR-UCB/EECS-2005-22 (November 2005)
5. Americans with Disabilities Act. (2010), <http://www.ada.gov>
6. Bahl, P., Padmanabhan, V.: RADAR: An In-Building RF-Based User Location and Tracking System. In: The Proceedings of IEEE Infocom., pp. 775–784 (2000)
7. Beckmann, C., Consolvo, S., LaMarca, A.: Some Assembly Required: Supporting End-User Sensor Installation in Domestic Ubiquitous Computing Environments. In: Davies, N., Mynatt, E.D., Siio, I. (eds.) *UbiComp 2004*. LNCS, vol. 3205, pp. 107–124. Springer, Heidelberg (2004)
8. Castro, P., Chiu, P., Kremenek, T., Muntz, R.R.: A Probabilistic Room Location Service for Wireless Networked Environments. In: Abowd, G.D., Brumitt, B., Shafer, S. (eds.) *UbiComp 2001*. LNCS, vol. 2201, pp. 18–34. Springer, Heidelberg (2001)
9. Crabtree, A., Rodden, T., Hemmings, T., Benford, S.: Finding a Place for UbiComp in the Home. In: Dey, A.K., Schmidt, A., McCarthy, J.F. (eds.) *UbiComp 2003*. LNCS, vol. 2864, pp. 208–226. Springer, Heidelberg (2003)
10. Cricket Series Mote. Crossbow Technologies, Inc. (2010), <http://www.xbow.com>
11. Edwards, W.K., Grinner, R.E.: At Home with Ubiquitous Computing: Seven Challenges. In: Abowd, G.D., Brumitt, B., Shafer, S. (eds.) *UbiComp 2001*. LNCS, vol. 2201, pp. 256–272. Springer, Heidelberg (2001)
12. Ekahau (2010), <http://www.ekahau.com>
13. Gray, D.: Mobility Impaired Individuals with Secondary Conditions: Health, Participation and Environments. In: The Final Report Submitted to Office of Disability and Health. CDC, Washington (2003)
14. Hazas, M., Ward, A.: A Novel Broadband Ultrasonic Location System. In: Borriello, G., Holmquist, L.E. (eds.) *UbiComp 2002*. LNCS, vol. 2498, pp. 264–280. Springer, Heidelberg (2002)
15. Hirsch, T., Forlizzi, J., Hyder, E., Goetz, J., Kurtz, C., Strobach, J.: The ELDer Project: Social, Emotional, and Environmental Factors in the Design of Eldercare Technologies. In: Proceedings of the ACM Conference on Universal Usability, pp. 72–79 (2000)

16. Indoor GPS (2010), <http://www.indoorgps.com>
17. Intille, S.S., Tapia, E.M., Rondoni, J., Beaudin, J.S., Kukla, C., Agarwal, S., Bao, L.: Tools for studying behavior and technology in natural settings. In: Dey, A.K., Schmidt, A., McCarthy, J.F. (eds.) UbiComp 2003. LNCS, vol. 2864, pp. 157–174. Springer, Heidelberg (2003)
18. Krumm, J., Cermak, G., Horvitz, E.: RightSPOT: A Novel Sense of Location for a Smart Personal Object. In: Dey, A.K., Schmidt, A., McCarthy, J.F. (eds.) UbiComp 2003. LNCS, vol. 2864, pp. 36–43. Springer, Heidelberg (2003)
19. LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, I., Sohn, T., Howard, J., Hughes, J., Potter, F., Tabert, J., Powledge, R., Borriello, G., Schilit, B.: Place Lab: Device Positioning Using Radio Beacons in the Wild. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) PERVASIVE 2005. LNCS, vol. 3468, pp. 116–133. Springer, Heidelberg (2005)
20. Madhavapeddy, A., Tse, T.: Study of Bluetooth Propagation Using Accurate Indoor Location Mapping. In: Beigl, M., Intille, S.S., Rekimoto, J., Tokuda, H. (eds.) UbiComp 2005. LNCS, vol. 3660, pp. 105–122. Springer, Heidelberg (2005)
21. New Freedom Initiative. Executive Order 13217 (2010),
<http://www.cms.hhs.gov/NewFreedomInitiative/>
22. O'Connell, T., Jensen, P., Dey, A.K., Abowd, G.D.: Location in the Aware Home. In: Workshop on Location Modeling for Ubiquitous Computing at Ubicomp 2001 (2001)
23. Otsason, V., Varshavsky, A., LaMarca, A., de Lara, E.: Accurate GSM Indoor Localization. In: Beigl, M., Intille, S.S., Rekimoto, J., Tokuda, H. (eds.) UbiComp 2005. LNCS, vol. 3660, pp. 141–158. Springer, Heidelberg (2005)
24. Patel, S.N., Kientz, J.A., Hayes, G.R., Bhat, S., Abowd, G.D.: Farther Than You May Think: An Empirical Investigation of the Proximity of Users to their Mobile Phones. In: Dourish, P., Friday, A. (eds.) UbiComp 2006. LNCS, vol. 4206, pp. 169–181. Springer, Heidelberg (2006)
25. Patel, S.N., Truong, K.N., Abowd, G.D.: PowerLine Positioning: A Practical Sub-Room-Level Indoor Location System for Domestic Use. In: Dourish, P., Friday, A. (eds.) UbiComp 2006. LNCS, vol. 4206, pp. 94–103. Springer, Heidelberg (2006)
26. Philipose, M., Fishkin, K.P., Perkowitz, M., Patterson, D.J., Fox, D., Kautz, H., Hahnel, D.: Inferring Activities from Interactions with Objects. IEEE Pervasive Computing 3(4), 50–57 (2004)
27. Priyantha, N.B., Chakraborty, A., Balakrishnan, H.: The Cricket Location-Support System. In: The Proceedings of The International Conference on Mobile Computing and Networking (Mobicom 2000), August 2000, pp. 32–43 (2000)
28. Rowan, J., Mynatt, E.D.: Digital Family Portrait Field Trial: Support for Aging in Place. In: The Proceedings of CHI 2005, pp. 521–530 (2005)
29. Tapia, E.M., Intille, S.S., Lopez, L., Larson, K.: The design of a portable kit of wireless sensors for naturalistic data collection. In: Fishkin, K.P., Schiele, B., Nixon, P., Quigley, A. (eds.) PERVASIVE 2006. LNCS, vol. 3968, pp. 117–134. Springer, Heidelberg (2006)
30. Ubisense (2010), <http://www.ubisense.net>
31. Varshavsky, A., Patel, S.N.: Location Systems. In: Krumm, J. (ed.) Ubiquitous Computing Fundamentals, pp. 286–319. CRC Press, Boca Raton (2009)
32. Vicon MX (2010), <http://www.vicon.com/products/system.html>
33. Want, R., Hopper, A., Falcao, V., Gibbons, J.: The active badge location system. ACM Transactions on Information Systems 10, 91–102 (1992)

Object-Based Activity Recognition with Heterogeneous Sensors on Wrist

Takuya Maekawa, Yutaka Yanagisawa, Yasue Kishino, Katsuhiko Ishiguro,
Koji Kamei, Yasushi Sakurai, and Takeshi Okadome

NTT Communication Science Laboratories

2-4 Hikaridai Seika-cho, Souraku-gun, Kyoto, Japan

{maekawa,yasue,ishiguro,yasushi}@cslab.kecl.ntt.co.jp,
yanagisawa.y@west.ntt.co.jp, kamei@atr.jp, tokadome@acm.org

Abstract. This paper describes how we recognize activities of daily living (ADLs) with our designed sensor device, which is equipped with heterogeneous sensors such as a camera, a microphone, and an accelerometer and attached to a user's wrist. Specifically, capturing a space around the user's hand by employing the camera on the wrist mounted device enables us to recognize ADLs that involve the manual use of objects such as making tea or coffee and watering plant. Existing wearable sensor devices equipped only with a microphone and an accelerometer cannot recognize these ADLs without object embedded sensors. We also propose an ADL recognition method that takes privacy issues into account because the camera and microphone can capture aspects of a user's private life. We confirmed experimentally that the incorporation of a camera could significantly improve the accuracy of ADL recognition.

Keywords: Wearable sensors; Recognizing daily activities; Experiment.

1 Introduction

Activity recognition is one of the most important tasks in pervasive computing applications. This task has a wide range of applications in, for example, context-aware systems, life logging and monitoring and has thus been the subject of a large amount of research. Two main approaches are used for activity recognition studies: environment augmentation and wearable sensing. The environment augmentation approach attempts to recognize users' activities by using sensors embedded in indoor environments. In the computer vision community, activity recognition tasks are accomplished by using cameras installed in a given environment. For example, hand washing and operating medical appliances can be recognized by domain specific solutions [23][28]. However, the task has become dominated by various types of embedded small sensors. Recently, many researchers in the field of ubiquitous computing have tried to recognize activities based on dense object usage sensors such as RFID tags and switch sensors installed in indoor environments [25][32][14]. With this approach, many studies recognize activities of daily living (ADLs) such as using the toilet, making coffee,

washing dishes, and taking medicine by using object usage sensors that are embedded in or attached to such daily use indoor objects and appliances as toilets, coffee makers, sinks, and cups.

The wearable sensing approach tries to recognize a user's activities by employing such sensors as body-worn accelerometers and microphones to capture characteristic repetitive motions, postures, and sounds of activities [19,20,23,16]. Using these types of wearable sensors, sensing studies have successfully recognized such activities as walking, bicycling, brushing teeth, speaking and laughing, and workshop activities such as sawing and drilling that have characteristic motions and/or sounds. An advantage of this approach is that it does not require environment embedded sensors. That is, this approach incurs no cost in terms of money or time for embedding sensors in indoor objects and furniture. Also, users can easily turn off their wearable devices when they want to preserve their privacy. The ADL recognition method proposed in this paper also uses body-worn sensors. However, because most existing studies use only such sensors as accelerometers and microphones, they cannot recognize ADLs that have no characteristic motions or sounds. For example, recognizing such ADLs as making tea and taking medicine, which the environment augmentation approach can achieve by using object usage sensors, is difficult when using only accelerometers and microphones. This study tries to recognize ADLs that involve object use by employing many kinds of sensors including cameras, microphones, and accelerometers attached to a single point on the body. In particular, to recognize these ADLs, we leverage visual features of objects, obtained from a camera on a user's wrist with which we may also easily capture such other features as the motion and sound of the ADLs. One of the characteristics of this study is that it incorporates the visual features of object use into wearable sensing. This permits us to recognize various kinds of ADLs that involve object use without the need for environment embedded sensors. To our knowledge, no work has reported object based ADL recognition employing the vision, sound, and motion features of object use captured by wrist worn sensors.

First, we describe the design of our proposed practical wearable sensor device, which is attached to one point on the body to recognize ADLs that involve object use, and then we build a prototype of the device. We report our design of a wristband type sensor device equipped with such sensors as a camera and a microphone. The device captures sensor data such as images of used objects and the sound emitted when a user performs an ADL and sends them wirelessly to a host PC. Second, we propose a supervised machine learning based ADL recognition method that uses the multi-modal sensor data. Note that, because the raw data obtained from a camera and a microphone on the user's wrist include private information, we design a recognition method where the sensor device does not send raw private information but abstracted information. Third, we collect sensor data by using the implemented prototype device. We capture ADLs that involve object use such as making tea, making green tea, taking medicine, vacuuming, washing dishes, and feeding fish, and annotate the collected data. Finally, we evaluate our recognition method by using the collected data and investigate



Fig. 1. (a) Conceptual image of wristband type sensor device and (b) prototype device

the contributions of each sensor. In summary, our contributions reported in this paper are (1) the design of a wearable device that enables us to recognize ADLs that involve object use without environment embedded sensors, (2) the proposal of an ADL recognition method that can detect ADLs involving object use, and (3) an experimental evaluation of the proposed method.

2 Practical Sensor Device

Our goal is to recognize ADLs that involve the use of objects. Designing a sensor device to achieve this goal, we must choose which types of sensor the device should be equipped with and select which point on the body the device should be attached to. We selected a camera, a microphone, an accelerometer, an illuminometer, and a digital compass from the range of commonly used sensors. We can expect both the cost and size of such sensors to decrease. Specifically, a camera captures visual information about objects used in ADLs. For example, an image (frame) including a coffee maker that is captured when a user makes coffee can be useful for recognizing the ADL of making coffee. The other four types of sensors are usually used for wearable activity recognition [19][16]. Also, we attach just one wristband type device equipped with the above five sensors to a user's dominant wrist. We attach the device to a single body location because wearing multiple devices on different parts of the body such as the waist, arms, and legs may place a large burden on the user in her daily life. Because almost all ADLs that involve object use are performed by hand, a sensor device attached near the hand can capture ADL characteristics well. Moreover, we can embed these sensors in a wristwatch.

Fig. 1(a) shows our ideal wristband sensor device designed based on the above discussion. We assume that the device sends preprocessed data obtained from the five sensors wirelessly to a host PC. Feature extraction and ADL recognition are performed on the PC (as shown in Fig. 4). The camera lens is placed on the inside of the wrist to capture the space around the wearer's hand because then the camera can capture objects held by the user and objects around her hand. Based on these assumptions, we fabricated the prototype wristband type sensor device shown in Fig. 1(b) for the experiment. We fixed together a USB camera, a wired microphone, and a USB cable wired sensor board with a 3-axis accelerometer, an illuminometer, and a 3-axis digital compass and attached them

to the wristband. The USB camera captures 352 by 288 pixel 24-bit color JPEG images at about 6 fps with an automatic focus and white balance function. We used a monaural omni-directional microphone with a sampling rate of 44.1 kHz. The sampling rates of the other three sensors on the sensor board were all about 30 Hz. The frequency of the accelerometer was sufficient compared with the 20 Hz frequency that is required to access daily activities [4]. We selected a small camera and a thin USB cable to avoid disturbing the user's activities. We also bound the sensor cables together with tape. The bundle was fixed in place with a band worn on the brachial region. These sensors are connected to a laptop carried in a backpack via the cables and they send their data to the laptop.

3 Proposed Method

We model each ADL class trained with annotated training data and use the models to classify test data. To recognize ADLs by using sensor data, training data should be acquired in each user's environment because these sensor data are environment dependent. For example, the sound of vacuuming may depend on the type of vacuum cleaner used in the environment. That is, users should label each ADL collected in their environment during a certain period of time. Models of ADL classes for ADL recognition are then generated by using features extracted from the annotated training data.

3.1 Annotating Training Data in Our Approach

To label an ADL, users should specify its ADL class and its start and end points. Our sensor device is equipped with a camera thus making it superior to those without a camera as regards labeling tasks. Assume that the sensor data are acquired from a sensor device with a microphone and an accelerometer on a certain day. After the data acquisition, it is almost impossible for users to annotate the acquired data solely by listening to the recorded sound. Here, we introduce two approaches that deal with the problem. The first is a method where users annotate the data while watching video recordings captured by cameras embedded in the environment [17]. However, it is very expensive to install cameras in various rooms in the users' houses to track their activities. The second approach uses an experience sampling method that permits users to make annotations in real time [12]. In one example, users carry a PDA that is used as a timing device to trigger self-reported activity entries. Although this approach is inexpensive, users have to be continuously aware of the annotation process. This may result in biased or unrealistic data [14]. To solve these problems, [14] proposes a voice based annotation method, which permits users to make annotations easily in real time via a headset. However, real time annotation methods have another problem in that users cannot easily modify mistakenly created labels. During long periods of training data acquisition, users are certain to produce incorrect labels. However, in an environment with no video recording, it is almost impossible for users to review them solely by referring to captured non-visual sensor data. In

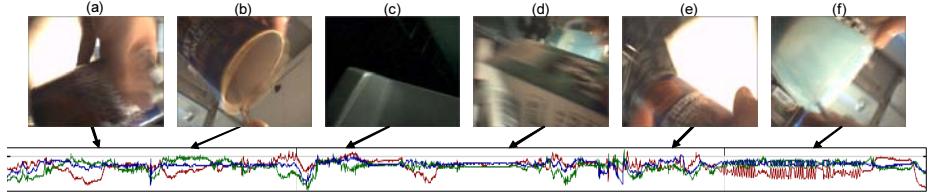


Fig. 2. Example camera and acceleration data for making cocoa

contrast, because our device has a camera, users can make an accurate label set by viewing image sequences recorded by the camera. Because the label set is used for training the models, it is very important to obtain training examples that are as accurate as possible [14].

As above, users annotate the sensor data while viewing an image sequence obtained by the camera and a chart of time-series acceleration data obtained from the device on our implemented annotation tool. By using the track bar of the tool, users can display an image captured at an arbitrary time on a panel component of the tool. Users can also play recorded image sequences and sound.

3.2 Classification Features

We extract features from annotated training data that are used to model and recognize ADL classes. We deal with time-series data obtained from various types of sensors with different sampling rates. Thus, after extracting features from the sensor data for each sensor type in an appropriate size window, we combine them into one second windows with a 50% overlap and compute averages for each feature in each window. The 50% overlap has been employed successfully in past studies [2]. We perform ADL modeling and recognition by using a feature vector sequence generated by combining features extracted from all the sensors. Here, we describe how to extract features from each sensor data.

Visual features. If we can detect which object the user is currently employing from the camera images, the information may be very useful for ADL recognition. In the following, after describing the characteristics of images captured by the camera and problems with the images such as privacy concerns, we use them as a basis for determining what kind of visual features are used to model ADL classes. Also, to achieve real time ADL recognition, we must extract the features from the image quickly. Note that we compute features for each captured image.

[Characteristics of camera images]

We introduce images captured by the camera in a data acquisition experiment. Fig. 2 shows a sequence of images and a chart of time-series acceleration data that were captured while a participant made cocoa. Fig. 2 (a) shows an image of when he took the cocoa tin from the cabinet, (b) shows an image of when he was spooning the cocoa powder, (c) shows an image of when he was moving toward the refrigerator, (d) shows an image of him holding a milk carton,

(e) shows an image of him holding the cocoa tin prior to storing it in the cabinet, and (f) shows an image of him stirring the cocoa. Images of objects captured by the wrist mounted camera have the following characteristics: (1) Objects are captured from various angles. (2) Most images show only a portion of the objects. (3) Objects seen in most images are blurred because of hand movement. (4) The brightness of an object can vary depending on the relationship between the lighting, camera, and object positions. Many studies try to detect objects from images while taking occlusion, rotation, scale, and blur into account [27][18]. However, to detect an object from images captured from various angles, we must generate a model of the object from many images of objects. This may place a large burden on the end user because we must generate expensive models for each end user environment. Also, most existing object recognition algorithms are very costly if they are designed to achieve real time ADL recognition. On the basis of the above, we consider that we can leverage only rough visual information.

[Problems with camera images]

We describe two problems related to images captured by the wrist camera. The first concerns privacy. We assume that the sensor device sends such sensor data wirelessly as camera images to a host PC. Users may feel reluctant to send images related to their private lives wirelessly, e.g., those captured in a toilet. The second problem relates to communication traffic. Continuously transmitting raw images in real time occupies a communication band constantly. Our implemented device requires about 90 KB/sec for raw image transmission. This may also exhaust the device batteries very quickly. As a result, we determined that the device should send images consisting of small quantities of abstracted data.

[Summary of our approach to visual feature extraction]

Based on the above, we decided to extract rough visual features from an abstracted image sent from the device. The data volume of this image is small and the image is secure. Specifically, we use a color histogram of an image sent from the device. Some studies also achieve fast object recognition/tracking [30][7] by comparing histograms and object models prepared in advance. In our approach, by using a histogram sent from the device, we simply count the number of pixels in the image (histogram) that are similar to a color characteristic of an ADL. For example, if a color of a cocoa tin is magenta, the number of pixels whose color is similar to magenta in an image may be useful for recognizing cocoa making activities. For each ADL, we obtain several characteristic colors from annotated training data in advance. For each characteristic color, we count the number of pixels in the histogram whose color is similar to the characteristic color. The result is used for the visual feature. Our purpose in using the histograms and characteristic colors is to achieve rough visual feature extraction with low communication and computation costs. In the following, we describe how to find the colors characteristic of each ADL, how to generate histograms, and how to compute features.

[Finding characteristic colors of each ADL]

We obtain the colors characteristic of each ADL in advance by using images of the annotated training data. Fig. 3(a) shows the procedure. (I) We cluster all the



Fig. 3. (a) Clustering pixels of all images of an ADL class and ranking the clusters (features) by their computed information gains, and (b) clustering color pixels of an image and constructing a histogram from the clusters

color pixels in all the raw images labeled as the ADL into 64 clusters by using the k-means algorithm in the hue, saturation, and brightness (HSB) color space with a slight modification. Then, we compute the average color of each cluster. This procedure provides 64 representative colors of the ADL. Here, we focus on the HSB color space because it has a brightness axis. As mentioned above, the brightness of an object can change depending on the positional relationship of the lighting, camera, and object. Thus, we multiply the brightness values of the pixels by 0.5 to reduce the importance of the brightness axis. (II) From the obtained 64 candidate (representative) colors of the ADL, we extract the top- m candidate colors as the characteristic colors of the ADL. We rank the 64 candidate colors in terms of information gain. The information gain is usually used to find distinguishable attributes (features) of instances. The information gain of an attribute increases the better the attribute classifies the instances. We compute each attribute's information gain when distinguishing images (instances) of the ADL class from those of other ADL classes by using the attribute values of the images. In this case, each attribute corresponds to the number of pixels in an image whose colors are similar to each candidate color. (How to count the similar pixels is mentioned below.) We compute the information gain of each attribute by using the computed attribute values of the images and then rank the attributes by their information gains to obtain the top- m attributes as characteristic colors of the ADL. Note that, before obtaining the top- m attributes (colors), we remove colors that are similar to other higher ranked colors from the ranking.

Here, we provide an example. Assume that the color of a cocoa tin used in making cocoa is magenta and other ADLs do not include objects whose colors are similar to magenta. The number of magenta pixels in an image captured while making cocoa is large and so the information gain of the attribute (the number of magenta pixels) becomes high because the attribute contributes to distinguish the cocoa making images from the others. (An image in which the number of magenta pixels is above a certain threshold may correspond to cocoa making images.) See [34] for detailed explanation of computing the information gain. From the above procedure, we can obtain m characteristic colors for each ADL.

[Histogram generation]

Fig. 3 (b) shows the procedure. (I) The device reduces the color of an image to 64 colors simply by using the k-means algorithm to cluster the pixels in the image into 64 clusters. The representative color of each cluster corresponds to an average

value for the colors in the cluster. (II) We compute a histogram of the image with 64 bins where each bin corresponds to one color of the 64 representative colors of the clusters. (The histogram is different from the commonly used color histogram.) The histogram includes only HSB data of each color (bin) and the number of pixels of the color in the reduced image. Comparing a characteristic color with colors of bins enables us to count the number of pixels in the image whose color is similar to the characteristic color. The histogram also permits us to compress an image into 64 pairs of 24 bit HSB data and 32 bit numeric data, i.e., $(24+32)*64 = 3584$ bit = 448 bytes. This enables us to reduce the communication traffic of our device, which captures images at 6 fps, to about 2.7 KB/sec. Moreover, we can solve the privacy problem because it is impossible to restore the original image from the histogram. Here, we use k-means clustering for color reduction in the device. We can process the algorithm at high speed by using a special purpose processing circuit [24]. We consider that all sensor data processing should be performed on special purpose circuits. Note that, in our prototype device shown in Fig. 1(b), the host PC performs the color reduction and histogram generation offline. Also, to annotate training data, our approach requires raw captured images as described above. The device should be designed to store raw images in its flash memory card during training data acquisition periods. This enables users to safely transmit the data to the PC via the card.

[Visual feature extraction]

For each characteristic color, on the host PC, we count the number of pixels in the histogram whose color is similar to the characteristic color to model and recognize ADLs. The similarity is computed by using the Euclidean distances between the colors in the modified HSB color space. That is, we simply count the number of pixels whose similarity is smaller than a threshold th . The approach is identical to that used for the characteristic color extraction. Then, we normalize the result by dividing it by the dimensions of the image. The normalized result corresponds to the visual feature. That is, the number of visual features extracted from one image corresponds to the number of characteristic colors.

We employ this simple method because it requires low computational power. In fact, this method can extract visual features from a histogram in about 0.5 msec on a PC with a 2.4 GHz CPU by using 75 characteristic colors. We set $m = 5$ and $th = 15$ because they resulted in good performance in a preliminary experiment.

Sound features. We extract features from sound that is emitted during ADLs that involve object use. For example, the sound of using a vacuum cleaner, tooth brushing, and running water may be useful for ADL recognition. We focus on the characteristic frequencies of such sounds. In [8], the Mel-Frequency Cepstral Coefficient (MFCC) is reported to be the best transformation scheme for environmental sound recognition. [5] achieves the highly accurate recognition of bathroom activities such as showering, flushing, and urination by using the MFCC. Thus, we decided to use the MFCC to recognize ADL related environmental sounds. Computing the MFCC is not expensive because it is based on Fast Fourier Transform (FFT). Note that sound recorded by the microphone

has problems related to data volume and privacy as well as the camera images. We thus extract sound features on the sensor device and send only them to the host PC. Also, the extraction of sound features from all sound data captured at a high sampling rate is costly. Thus, we intermittently capture short periods of sound and then compute a 13 order MFCC of each captured sound windowed by a Hamming window. In this implementation, we record 25 milliseconds of sound six times a second. From the sound data, we can obtain twelve features.

Acceleration features. We can extract features of postures and repetitive hand movements from acceleration data. For example, we can find a characteristic frequency in the acceleration data captured while the participant was stirring cocoa as shown in Fig. 2 (f). We extract features based on the FFT components of each 64 sample window acceleration data. We use the mean, energy, frequency-domain entropy, and dominant frequency as features. The mean can characterize the hand posture. For example, the hand posture during tooth brushing may have particular characteristics. The mean is the DC component of the FFT. The energy can be used to distinguish low intensity activities such as standing from high intensity activities such as walking [33]. The energy feature is calculated by summing the magnitudes of the squared discrete FFT components. For normalization, the sum was divided by the window length. Note that the DC component of the FFT is excluded from this summation. The frequency-domain entropy and dominant frequency can distinguish between repetitive motions with similar energy values. For example, the major FFT frequency components of stirring cocoa were between about 2 and 4 Hz in our experiment. Those of brushing teeth were between about 4 and 6 Hz. The frequency-domain entropy is calculated as the normalized information entropy of the discrete FFT component magnitudes [2]. The dominant frequency is the frequency that has the largest FFT component, and this component is three times larger than the average component of all the frequencies in this implementation. If there is no frequency that satisfies the conditions, we set the feature at zero. As above, we extract a total of twelve features from the 3-axis acceleration data.

Illuminance and direction features. We use raw sensor data captured by the illuminometer and 3-axis digital compass directly as features. The digital compass captures the characteristic human orientation of each ADL. Assume that a user habitually brushes her teeth in front of a sink in her house. Her orientation during brushing may always be the same.

3.3 Classification Methodology

We model ADL classes in advance by using annotated training data and the obtained feature vector sequence. We classify each feature vector in the test data into an estimated ADL class. The classification approaches used in machine learning are divided into two groups: one group uses discriminative techniques that learn the class boundaries and the other uses generative techniques that model the conditional density functions of the data classes. The classification

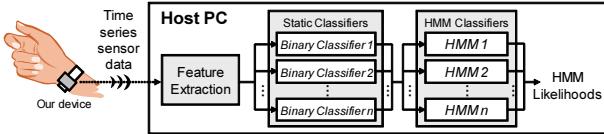


Fig. 4. Overview of the classification method

performance of the discriminative techniques, which find discriminant features of the classes, often outperform those of generative techniques. By contrast, handling missing data is often easier with the generative techniques. The ML community has shown increasing interest in a hybrid discriminative/generative approach that can combine the advantages of the two techniques [13][26]. State-of-art activity recognition studies also achieve high accuracy by employing this approach [15][16][11]. In addition, because we deal with time-series data, incorporating the hidden Markov model (HMM), which is a generative model that can be used to model activities with temporal patterns [20][14], into the hybrid approach, can improve the performance and smoothness of ADL recognition.

These facts provide our motivation for using the hybrid discriminative/generative approach with HMMs. Hybrid classification employs two main modules: static classifiers and HMM classifiers as shown in Fig. 4. The input of the first module is the extracted feature vector sequence. The first module consists of some discriminative binary classifiers trained with feature vector sequences. We build each binary classifier to recognize its corresponding ADL class. That is, the number of binary classifiers n corresponds to the number of ADLs the method learns. Each binary classifier computes the class probability for each feature vector in the feature vector sequence. That is, each binary classifier outputs the class probability sequence. The input of the second module consists of the class probability sequences computed by the n binary classifiers. The second module also comprises some HMM classifiers trained with a sequence of output class probabilities of the static classifiers. We also build each HMM to recognize its corresponding ADL class, that is, each HMM also outputs the likelihood of its corresponding ADL. The class with the highest likelihood is the classified class. We train the HMMs using the class probabilities of the static classifiers, which provide high levels of performance. The use of HMMs can smooth out sporadic errors of the static classifiers.

4 Data Collection

For our experimental evaluation, we collected sensor data from participants by using our prototype device shown in Fig. II(b). Then, each participant annotated her own data using our annotation tool. In this study we learned the fifteen ADLs listed in Table II. We selected these ADLs, which involve daily objects, by referring to ADLs dealt with by some reported ADL recognition studies.

Table 1. ADLs and their average duration (min)

	ADL	duration (min)		ADL	duration (min)
A	brush teeth	3.65	I	make juice	1.77
B	cook pasta	5.98	J	make tea	1.37
C	cook rice	4.33	K	practice aromatherapy	0.66
D	feed fish	0.40	L	take supplement	0.82
E	listen to music	1.69	M	vacuum	1.26
F	make cocoa	1.37	N	wash dishes	3.68
G	make coffee	1.63	O	water plants	0.27
H	make green tea	1.16			

4.1 Data Set

The most natural data would be acquired from the normal daily lives of the participants. However, obtaining sufficient samples of such data is costly because researchers have to observe their normal daily lives. We collect sensor data by using a semi-naturalistic collection protocol [2] that permits greater variability in participant behavior than laboratory data. In the protocol, participants perform a random sequence of ADLs (obstacles) following instructions on a worksheet. The participants are relatively free as regards how they perform each ADL because the instructions on the the worksheet are not very strict, e.g., “vacuum the room” and “listen to an arbitrary track from a CD in the rack.”

Data were collected from 10 participants who wore our prototype device in our experimental environments. The participants were workers (not researchers) in our laboratory. Because the features of the sensor data obtained from our device may vary depending on the environment, we collected sensor data in two environments: environment 1 and environment 2, and tested our method by using each data set. That is, we evaluated the test data obtained in environment 1 by using a classifier trained on training data also obtained in environment 1 and vice versa. Environment 1 is our home-like experimental environment [21]. We had equipped the environment with a cabinet, break time items, cooking utensils, etc. to emulate a home environment. In the experiment, we used objects originally installed in the environment. Also, four video cameras were fixed to the ceiling of the environment. The participants were familiar with environment 1 because they entered and left it many times every day. Because environment 2 is simply a room in our laboratory, we equipped it with new objects required for the ADLs for our experiment. We taught the participants the locations of the objects before undertaking data collection. Each data collection session included a random sequence of the fifteen ADLs listed in Table 1. We conducted fourteen sessions, which correspond to about two weeks data, in each environment. That is, each participant took part in a total of about three sessions in two environments. When performing the ‘brush teeth,’ ‘cook rice,’ and ‘wash dishes’ ADLs, they used sinks outside the environments. During data collection in environment 2, the participants used a timing device. The timing device is a PDA and can record the time at which its button is pushed. The participants can easily annotate collected data by referring to the recorded times. For example, they can push the button just before starting to vacuum.

The data obtained in this experiment were various and practical. Because the experiments were conducted from 9 a.m. to 6 p.m., images obtained under various light conditions are included. Also, because the experiment involved ten participants, their ways of performing the ADLs differed. For example, some participants made tea while standing and others while sitting. Of course, the participants' clothes, which were sometimes captured by the camera, were also different in different sessions. Furthermore, the experiment involved various kinds of objects such as those with complex textures, e.g., floral and arabesque patterns and translucent objects. Also, colors of some objects were similar with each other.

4.2 Labeling Sensor Data

The participants annotated their own collected data by using our tool. We asked them to select the start and end points of labels as they liked. After they had completed the task, we asked them to provide comments about the tasks. To enable us to compare our annotation method with conventional labeling in laboratory settings, the participants also annotated their data for environment 1 by watching video recordings captured by the cameras fixed to the ceiling. We call the label sets of environments 1 and 2 obtained by using the sensor data provided by our device *label sets 1A and 2*. We call the label set of environment 1 obtained using the video recordings provided by the fixed cameras *label set 1B*.

The average times needed to label the sensor data for one session were 44.1 and 36.4 min for set 1A and set 2, respectively. The participants annotated the data of environment 2 while referring to a printed list of times recorded by the timing device. Although we found no significant difference between two sets of results with a two-tail t-test ($p > 0.05$), all the participants commented that the timing device was useful. When end users annotate sensor data obtained during training data acquisition periods in their daily lives, they should determine their ADLs from sensor data obtained over long periods. Thus, the timing device may be useful to end users. The timing function should be embedded in our wristband device. The average labeling time for label set 1B was 27.0 min. While this approach is very costly, the average time was shorter than that of our device. Also, some participants commented that the images captured with our device can cause motion sickness. However, they also commented that labeling by using the images provided by our device was easier than they had thought because they could easily recognize routinely used objects in the images.

In label set 1A, there were three incorrect labels: a 'cook pasta' label did not include about half of a boil water activity in the ADL, a 'make juice' label ended while the participant was using a juicer, and a 'vacuum' label included part of another ADL. In label set 2, a participant forgot to label a 'feed fish' ADL. In label set 1B, there were two incorrect labels: a 'cook pasta' label did not include about half of a boil water activity and a 'listen to music' label did not cover the whole ADL. We could not find any significant differences between label sets 1A and 1B in terms of labeling accuracy. We asked the participants to correct these mistakes. In addition, each participant had a different labeling strategy. When labeling 'brush teeth' and 'wash dishes,' six participants selected

start and end points to include walking with related objects such as a dish rack from the environment to the sink. The labels of four other participants did not include this. In addition, some labels of the two participants did not include the ADL preparation time. For example, when making cocoa, the participants have to prepare a cup, a cocoa tin, and a milk carton. We should instruct end users in the same environment to establish a consensus on labeling strategy. We asked the participants in the minority to modify their labels in accordance with those of the majority.

5 Evaluation

We evaluated the performance of our method by using the annotated sensor data (label sets 1A and 2). We conducted a ‘leave-one-session-out’ cross validation evaluation. That is, we tested one session by using a classifier trained on thirteen other sessions. In this evaluation experiment, we used AdaBoost M1 and the C4.5 decision tree implemented on the Weka toolkit [34] as binary classifiers. AdaBoost is a boosting algorithm that combines weak classifiers to construct a strong classifier. We use a decision stump as a weak classifier.

5.1 Performance of Our Method

Table 2 lists the accuracies of the various recognition methods in some metrics. The AdaBoost+HMM (window) and C4.5+HMM (window) columns present precisions and recalls calculated based on feature windows (vectors). That is, precision is the ratio of the number of feature windows correctly classified into an ADL class to the number of all feature windows classified into the class. Recall is the ratio of the number of feature windows correctly classified into an ADL class to the number of actual feature windows of the class. C4.5+HMM, which uses C4.5 as a discriminative binary classifier and HMMs as a generative classifier, achieves relatively high accuracies for many ADLs and outperforms AdaBoost+HMM, which uses AdaBoost and HMMs. The accuracies of AdaBoost+HMM for certain ADLs such as ‘feed fish,’ ‘take supplement,’ and ‘water plants’ whose duration was short were zero. Because of the short duration of these ADLs, few feature windows were labeled as these ADLs. The AdaBoost algorithm combines weak classifiers, which usually ignore a minor class because they can achieve high accuracy by classifying all instances (windows) into a major class. This led to zero accuracies for these ADLs. [16] achieved the highly accurate recognition of primitive activities such as walking and sitting with a combination of AdaBoost and HMM. However, it was difficult to use this combination to recognize complex and/or brief ADLs.

The accuracies of C4.5+HMM for short duration ADLs such as ‘feed fish,’ ‘practice aromatherapy,’ ‘take supplement,’ and ‘water plants’ were also relatively low. This was caused by the head and foot margins of the labels. Hand-crafted labels inevitably start and end with margins with no distinguishable feature window. For example, in ‘take supplement,’ the margin can correspond

Table 2. Averaged accuracies (precision / recall) of the recognition methods. The values are percentages.

	AdaBoost+HMM (window)		C4.5+HMM (window)		AdaBoost+HMM (instance)		C4.5+HMM (instance)	
	Env. 1	Env. 2	Env. 1	Env. 2	Env. 1	Env. 2	Env. 1	Env. 2
A: brush teeth	42.1/73.0	75.2/91.4	74.3/79.0	84.3/88.1	27.5/78.6	50.0/92.9	92.9/92.9	77.8/100
B: cook pasta	97.3/86.4	99.2/90.4	97.2/83.7	98.7/84.7	100/92.9	100/100	100/100	100/92.9
C: cook rice	76.2/93.1	79.1/96.0	88.3/85.1	88.3/87.5	54.2/92.9	66.7/100	81.2/92.9	87.5/100
D: feed fish	44.9/3.0	0.0/0.0	60.5/67.7	74.1/58.7	0.0/0.0	0.0/0.0	92.3/85.7	88.9/57.1
E: listen to music	86.7/81.2	50.2/65.3	84.7/90.1	58.4/82.4	80.0/85.7	45.0/64.3	93.3/100	72.2/92.9
F: make cocoa	0.0/0.0	87.9/72.0	74.6/64.4	85.2/76.4	0.0/0.0	84.6/78.6	91.7/78.6	92.9/92.9
G: make coffee	36.4/61.3	49.2/77.8	73.8/66.5	85.2/90.4	24.2/57.1	40.7/78.6	69.2/64.3	93.3/100
H: make green tea	16.4/16.6	69.9/7.0	50.1/13.8	34.5/72.9	18.8/21.4	100/7.1	40.0/14.3	45.8/84.6
I: make juice	86.1/72.9	27.0/53.1	79.7/78.2	76.4/70.4	92.3/85.7	17.9/50.0	93.3/100	92.3/85.7
J: make tea	0.0/0.0	72.1/47.8	24.5/70.3	72.7/42.3	0.0/0.0	60.0/42.9	47.6/71.4	75.0/42.9
K: practice aroma.	66.2/38.7	97.4/57.7	72.8/68.6	77.1/75.4	83.3/35.7	90.9/71.4	100/85.7	100/85.7
L: take supplement	0.0/0.0	0.0/0.0	50.8/69.2	73.7/62.4	0.0/0.0	0.0/0.0	70.6/85.7	90.9/71.4
M: vacuum	96.8/82.0	89.4/80.1	89.0/87.8	93.2/83.1	100/85.7	86.7/92.9	100/100	100/92.9
N: wash dishes	98.3/80.9	97.6/77.5	93.1/82.6	94.3/89.9	100/85.7	100/92.9	93.3/100	93.3/100
O: water plants	100/88.4	0.0/0.0	84.5/92.4	40.5/59.8	100/100	0.0/0.0	100/100	100/71.4
Average	56.5/51.8	59.6/54.4	73.2/73.3	75.8/75.0	52.0/54.8	56.2/58.1	84.4/84.8	87.3/84.7

to a time duration where a participant walks from a chair to a cabinet to get a pill case. Feature windows involved in these margins may be wrongly classified. For an ADL with a short duration, the ratio of the time duration of its margins to those of the whole label is large and so the accuracy becomes relatively low. However, in C4.5+HMM, most feature windows in each label were correctly classified. For ease of understanding, the AdaBoost+HMM (instance) and C4.5+HMM (instance) columns in Table 2 show instance based accuracies, which are computed based on majority voting. That is, we compute the accuracies based on the strategy: In an ADL instance, we regard the instance itself to be classified into the majority vote of the recognition results of each feature window included in the instance. While our recognition method depends on environmental conditions, C4.5+HMM achieved high accuracies in both environments.

Distinguishing between ‘make green tea’ and ‘make tea’ was difficult in both environments as also described in Table 3, which shows the confusion matrices of C4.5+HMM in environments 1 and 2. This is because the motions involved in making green tea and making tea are the same, and most of the objects used in these ADLs such as a kettle and a cup are also the same. In addition, each side of the tea caddy in environment 1 is a single color; red or gold. In many sessions, the camera on our device could capture only the red colored portion of the caddy depending on how the caddy was held. Because the green tea caddy is also red, it was difficult to distinguish these ADLs in environment 1. Also, the recognition of such ADLs as ‘feed fish,’ ‘take supplement,’ and ‘water plants,’ which involve small numbers of objects and do not have characteristic sound or hand activities sometimes failed. In particular, when the colors of objects involved in the ADLs were similar to those of objects used in other ADLs, it was difficult to distinguish between these ADLs. In environment 2, for example, the color of the fish food tin was similar to that of a kettle used for making tea.

Table 3. Instance based confusion matrices of C4.5+HMM in environments 1 and 2

		Env. 1														Env. 2													
		A: brush teeth							B: cook pasta							A: brush teeth							B: cook pasta						
		C: cook rice	D: feed fish	E: listen to music	F: make cocoa	G: make coffee	H: make green tea	I: make juice	J: make tea	K: practice aroma.	L: take supplement	M: vacuum	N: wash dishes	O: water plants	C: cook rice	D: feed fish	E: listen to music	F: make cocoa	G: make coffee	H: make green tea	I: make juice	J: make tea	K: practice aroma.	L: take supplement	M: vacuum	N: wash dishes	O: water plants		
A	13	0	0	0	0	0	0	0	0	0	0	0	1	0	A	14	0	0	0	0	0	0	0	0	0	0	0	0	
B	0	14	0	0	0	0	0	0	0	0	0	0	0	0	B	0	13	0	0	0	0	0	0	0	0	0	0	0	
C	1	0	13	0	0	0	0	0	0	0	0	0	0	0	C	0	0	14	0	0	0	1	0	0	0	0	0	0	
D	0	0	1	12	0	0	0	0	0	0	0	0	0	0	D	1	0	0	8	2	0	0	0	3	0	0	0	0	
E	0	0	0	0	14	0	0	0	0	0	0	0	0	0	E	0	0	1	0	13	0	0	0	0	0	0	0	0	
F	0	0	0	0	0	11	1	0	1	0	0	1	0	0	F	0	0	0	0	0	13	0	1	0	0	0	0	0	
G	0	0	0	0	1	0	0	9	1	0	1	0	2	0	G	0	0	0	0	0	14	0	1	0	0	0	0	0	
H	0	0	0	0	0	0	0	2	2	0	9	0	1	0	H	0	0	0	0	0	0	11	0	2	0	0	0	0	
I	0	0	0	0	0	0	0	0	14	0	0	0	0	0	I	1	0	0	0	1	0	0	0	12	0	0	0	0	
J	0	0	0	0	0	1	0	2	0	10	0	1	0	0	J	0	0	0	0	0	1	1	5	0	6	0	1	0	
K	0	0	1	0	1	0	0	0	0	0	12	0	0	0	K	1	0	0	0	1	0	0	0	0	0	12	0	0	
L	0	0	1	0	0	0	0	0	0	1	0	12	0	0	L	1	0	0	0	0	0	0	2	1	0	0	10	0	
M	0	0	0	0	0	0	0	0	0	0	0	0	14	0	M	0	0	0	0	0	0	0	0	0	0	13	1	0	
N	0	0	0	0	0	0	0	0	0	0	0	0	0	14	N	0	0	0	0	0	0	0	0	0	0	0	14	0	
O	0	0	0	0	0	0	0	0	0	0	0	0	0	14	O	0	0	1	1	0	0	0	2	0	0	0	0	10	

5.2 Contributions of Each Sensor

Table 4(a) lists the accuracies of the C4.5+HMM recognition results in various sensor combinations. For example, the ‘only camera’ row shows instance based accuracies under a condition where the accuracies were computed on the basis of only features extracted from the camera sensor data. Also, the ‘w/o camera’ row shows accuracies under a condition where the accuracies were computed without features extracted from the camera sensor data. Surprisingly, we could achieve very high accuracies (about 75%) with just the camera. We also found that using only a camera could achieve almost the same accuracies as when combining an accelerometer, a microphone, a direction sensor, and an illuminometer. The camera played a significant role in ADL recognition when using our device. Sensors with a high contribution were the camera, accelerometer, and microphone in that order. The illuminometer and digital compass barely contributed to the recognition and sometimes even decreased the accuracy.

Table 4 (b) lists the accuracies of each ADL when we use only the camera and only the accelerometer. The accuracies of most ADLs when using only the camera were high. However, it was difficult to distinguish between ‘make tea’ and ‘make green tea’ in environment 1 because the colors of the objects involved in these ADLs were similar. Also, the accuracies for ‘cook pasta’ and ‘listen to music,’ which were characterized by their sound features, were not very high. With only the accelerometer, the accuracies of ‘brush teeth,’ ‘cook rice,’ and ‘wash dishes’ were relatively high. However, without the camera, it was difficult to distinguish between these ADLs with high accuracy because all three ADLs, which involved long periods of walking (and the sound of running water), were similar. Moreover, without a camera, it is very difficult to distinguish such ADLs

Table 4. (a) instance based average accuracies (precision/recall) of C4.5+HMM in various sensor combinations and (b) instance based average accuracies of C4.5+HMM for each ADL with only camera features and with only accelerometer features

(a)				(b)			
Sensor	Condition	Env.	Accuracy		only camera		only accelerometer
	only	1	76.7/73.2		Env. 1	Env. 2	Env. 1
camera	only	2	75.1/71.8	A: brush teeth	75.0/85.7	41.2/50.0	71.4/71.4
		1	77.7/75.2	B: cook pasta	88.9/57.1	31.2/35.7	39.3/78.6
	w/o	2	71.8/67.6	C: cook rice	70.0/100	72.2/92.9	50.0/78.6
		1	28.3/32.9	D: feed fish	90.9/76.9	88.9/57.1	100/7.1
micro-phone	only	2	21.8/28.6	E: listen to music	72.2/92.9	62.5/71.4	53.8/50.0
		1	84.9/83.3	F: make cocoa	66.7/42.9	84.6/78.6	23.5/28.6
	w/o	2	83.8/81.0	G: make coffee	84.6/78.6	73.3/78.6	61.5/57.1
		1	48.5/44.3	H: make green tea	35.0/50.0	52.4/84.6	7.7/7.1
accelerometer	only	2	47.3/43.8	I: make juice	76.5/92.9	81.8/64.3	100/78.6
		1	82.1/80.5	J: make tea	27.8/35.7	83.3/71.4	9.5/14.3
	w/o	2	84.9/79.5	K: practice aroma.	100/73.4	100/92.3	0.0/0.0
		1	0.1/6.7	L: take supplement	77.8/50.0	81.8/64.3	0.0/0.0
illuminometer	only	2	0.4/6.7	M: vacuum	100/78.6	85.7/85.7	86.7/92.9
		1	81.7/82.4	N: wash dishes	85.7/85.7	87.5/100	66.7/71.4
	w/o	2	89.6/88.0	O: water plants	100/100	100/50.0	100/42.9
		1	23.1/21.9	Average	76.7/73.2	75.1/71.8	48.5/44.3
digital compass	only	2	10.8/10.0				47.3/43.8
		1	85.9/84.8				
	w/o	2	89.9/87.0				

as ‘feed fish,’ ‘practice aromatherapy,’ ‘take supplement,’ and ‘water plants.’ These ADLs have few distinguishable features other than visual features. We consider that, without the camera, it is difficult to recognize the complex ADLs studied here.

From the above results, we consider that the wrist is a good place on the body to attach a single sensor device designed to capture ADLs that involve object use. Hand posture (mean) contributed to the recognition of many ADLs such as ‘brush teeth’ and ‘make juice’. However, it is difficult to capture the features when using body locations other than the wrist. Without the mean features, instance based precision and recall decreased to 82.3 and 79.5 in environment 1. In addition, the wrist worn camera, which was the best contributor, can easily capture hand manipulated objects. [22] uses a shoulder mounted robot with a camera to recognize and record hand activities such as operating a keyboard and operating a calculator. However, the robot has to control the direction of its camera to track the hand.

6 Related Work

We introduce related work that relates to vision based wearable sensing. [9] achieves gait analysis and floor recognition by using a shoe mounted camera and accelerometers. Floor recognition permits us to know the user’s location. [35] recognizes kitchen activities by using RFID tags attached to kitchen objects and a camera that overlooks the kitchen counter. An RFID reader bracelet worn on a user’s wrist and the camera detect the use of the objects. [6] uses a camera and a microphone attached to a chest strap to detect location related events such as

entering an office, kitchen, or courtyard. [29] uses two hat mounted cameras to determine user's actions in a game, e.g., aiming a gun. On the other hand, we use a wrist mounted camera, a microphone, an accelerometer, an illuminometer, and a digital compass to recognize ADLs that involve object use by capturing various characteristic features of manually used objects. Also, [1] uses a wrist mounted camera to realize a virtual keyboard by tracking the fingers with the camera.

There exists some works that recognize activities by using a single sensor device embedded in a home. These works also attempt to reduce costs of sensor deployment. For example, HydroSense [10] employs a water pressure sensor to understand activities that involve water use.

7 Conclusion and Future Work

We implemented a prototype wristband sensor device to recognize ADLs that involve the manual use of objects. The device is equipped with a camera, a microphone, an accelerometer, an illuminometer, and a digital compass to capture various characteristic features of object use. This device enables us to recognize various kinds of ADLs that existing wearable sensor devices cannot recognize without environment embedded sensors. In the experiments, we confirmed that the incorporation of a camera could achieve highly accurate ADL recognition.

As a part of our future work, we plan to solve the problems thrown up by the experiment. In both the environments, it was difficult to distinguish between such ADLs as 'make green tea' from 'make tea' that involve the same hand activities and the similar colored objects. To cope with such scalability problems, we should extract more detailed features such as SIFT features [18] from 'good' images, e.g., those including logos, while taking account of privacy concerns and communication costs. Furthermore, our ML-based approach cannot deal with situations where residents replace objects, e.g., residents frequently replace milk cartons. Because the types of milk that a family regularly purchases may be limited, we should instruct end users to prepare ADL training data that include various product types of such objects or we should realize an object replacement detection technique to induce users to prepare new training data.

We also plan to develop a new wristband sensor device that works without a laptop. The device permits us to capture sensor data in real environments and evaluate the performance of our method by using the data.

Acknowledgments. The authors would like to thank Dr. Takuya Yoshioka for the helpful comments and discussions.

References

1. Ahmad, F., Musilek, P.: A keystroke and pointer control input interface for wearable computers. In: Proc. PerCom 2006, pp. 2–11 (2006)
2. Bao, L., Intille, S.S.: Activity recognition from user-annotated acceleration data. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 1–17. Springer, Heidelberg (2004)

3. Blum, M., Pentland, A.S., Troster, G.: Insense: Interest-based life logging. *IEEE Multimedia* 13(4), 40–48 (2006)
4. Bouten, C.V., et al.: A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity. *IEEE Trans. on Bio-Medical Engineering* 44(3), 136–147 (1997)
5. Chen, J., Kam, A.H., Zhang, J., Liu, N., Shue, L.: Bathroom activity monitoring based on sound. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) *PERVASIVE 2005*. LNCS, vol. 3468, pp. 47–61. Springer, Heidelberg (2005)
6. Clarkson, B., Mase, K., Pentland, A.: Recognizing user context via wearable sensors. In: Proc. ISWC 2000, pp. 69–75 (2000)
7. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. *IEEE Trans. on Pattern Analysis Machine Intelligence* 25(5), 564–577 (2003)
8. Cowling, M.: Non-speech environmental sound recognition system for autonomous surveillance. Ph.D. Thesis, Griffith University, Gold Coast Campus (2004)
9. Fitzpatrick, P., Kemp, C.C.: Shoes as a platform for vision. In: Proc. ISWC 2003, pp. 231–234 (2003)
10. Froehlich, J.E., Larson, E., Campbell, T., Haggerty, C., Fogarty, J., Patel, S.N.: HydroSense: Infrastructure-mediated single-point sensing of whole-home water activity. In: Proc. Ubicomp 2009, pp. 235–244 (2009)
11. Huynh, T., Schiele, B.: Towards less supervision in activity recognition from wearable sensors. In: Proc. ISWC 2006, pp. 3–10 (2006)
12. Intille, S.S., Tapia, E.M., Rondoni, J., Beaudin, J., Kukla, C., Agarwal, S., Bao, L., Larson, K.: Tools for studying behavior and technology in natural settings. In: Dey, A.K., Schmidt, A., McCarthy, J.F. (eds.) *UbiComp 2003*. LNCS, vol. 2864, pp. 157–174. Springer, Heidelberg (2003)
13. Jaakkola, T., Haussler, D.: Exploiting generative models in discriminative classifiers. In: Proc. Advances in Neural Information Processing Systems, vol. 11, pp. 487–493 (1999)
14. Kasteren, T.V., Noulas, A., Englebienne, G., Kroese, B.: Accurate activity recognition in a home setting. In: Proc. UbiComp 2008, pp. 1–9 (2008)
15. Lester, J., Choudhury, T., Kern, N., Borriello, G., Hannaford, B.: A hybrid discriminative/generative approach for modeling human activities. In: Proc. IJCAI 2005, pp. 766–772 (2005)
16. Lester, J., Choudhury, T., Borriello, G.: A practical approach to recognizing physical activities. In: Fishkin, K.P., Schiele, B., Nixon, P., Quigley, A. (eds.) *PERVASIVE 2006*. LNCS, vol. 3968, pp. 1–16. Springer, Heidelberg (2006)
17. Logan, B., Healey, J., Philipose, M., Tapia, E.M., Intille, S.S.: A long-term evaluation of sensing modalities for activity recognition. In: Krumm, J., Abowd, G.D., Seneviratne, A., Strang, T. (eds.) *UbiComp 2007*. LNCS, vol. 4717, pp. 483–500. Springer, Heidelberg (2007)
18. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int'l Journal on Computer Vision* 60(2), 91–110 (2004)
19. Lukowicz, P., Junker, H., et al.: WearNET: a distributed multi-sensor system for context aware wearables. In: Borriello, G., Holmquist, L.E. (eds.) *UbiComp 2002*. LNCS, vol. 2498, pp. 361–370. Springer, Heidelberg (2002)
20. Lukowicz, P., Ward, J., Junker, H., Stager, M., Troster, G., Atrash, A., Starner, T.: Recognizing workshop activity using body worn microphones and accelerometers. In: Ferscha, A., Mattern, F. (eds.) *PERVASIVE 2004*. LNCS, vol. 3001, pp. 18–32. Springer, Heidelberg (2004)

21. Maekawa, T., Yanagisawa, Y., Kishino, Y., Kamei, K., Sakurai, Y., Okadome, T.: Object-blog system for environment-generated content. *IEEE Pervasive Computing* 7(4), 20–27 (2008)
22. Mayol, W.W., Murray, D.W.: Wearable hand activity recognition for event summarization. In: Proc. ISWC 2005, pp. 122–129 (2005)
23. Mihailidis, A., Carmichael, B., Boger, J.: The use of computer vision in an intelligent environment to support aging-in-place, safety, and independence in the home. *IEEE Trans. on Info. Tech. in BioMedicine* 8(3), 238–247 (2004)
24. Morikawa, S., Ito, K., Shibata, T.: A k-means VLSI processor and its application to autonomous area segmentation in images. *IEIC Technical Report* 106(342), 19–24 (2006)
25. Philipose, M., Fishkin, K.P., Perkowitz, M.: Inferring activities from interactions with objects. *IEEE Pervasive Computing* 3(4), 50–57 (2004)
26. Raina, R., Shen, Y., Ng, A.Y., McCallum, A.: Classification with hybrid generative/discriminative models. In: Proc. Advances in Neural Information Processing Systems, vol. 16 (2003)
27. Schiele, B., James, L.C.: Object recognition using multidimensional receptive field histograms. In: Proc. European Conference on Computer Vision, pp. 610–619 (1996)
28. Shi, Y., Huang, Y., Minnen, D., Bobick, A., Essa, I.: Propagation networks for recognition of partially ordered sequential action. In: Proc. CVPR 2004, vol. 2, pp. 862–869 (2004)
29. Starner, T., Schiele, B., Pentland, A.: Visual contextual awareness in wearable computing. In: Proc. ISWC 1998, pp. 50–57 (1998)
30. Swain, M.J., Ballard, D.H.: Color indexing. *Int'l Journal of Computer Vision* 7, 11–32 (1991)
31. Tapia, E.M., Intille, S.S., Larson, K.: Activity recognition in the home using simple and ubiquitous sensors. In: Ferscha, A., Mattern, F. (eds.) *PERVASIVE 2004. LNCS*, vol. 3001, pp. 158–175. Springer, Heidelberg (2004)
32. Tapia, E.M., Intille, S.S., Larson, K.: Portable wireless sensors for object usage sensing in the home: challenges and practicalities. In: Schiele, B., Dey, A.K., Gellersen, H., de Ruyter, B., Tscheligi, M., Wichert, R., Aarts, E., Buchmann, A. (eds.) *AmI 2007. LNCS*, vol. 4794, pp. 19–37. Springer, Heidelberg (2007)
33. Welk, G., Differding, J.: The utility of the Digi-Walker step counter to assess daily physical activity patterns. *Medicine & Science in Sports & Exercise* 32(9), S481–S488 (2000)
34. Witten, I.H., Frank, E.: *Data Mining: Practical machine learning tools and techniques*, 2nd edn. Morgan Kaufmann, San Francisco (2005)
35. Wu, J., Osuntogun, A., et al.: A scalable approach to activity recognition based on object use. In: Proc. ICCV 2007, pp. 1–8 (2007)

GasSense: Appliance-Level, Single-Point Sensing of Gas Activity in the Home

Gabe Cohn¹, Sidhant Gupta², Jon Froehlich², Eric Larson¹, and Shwetak N. Patel^{1,2}

¹ Electrical Engineering, ² Computer Science & Engineering

UbiComp Lab, DUB Group, University of Washington
Seattle, WA, 98195

{gabecohn, sidhant, jfroehli, eclarson, shwetak}@uw.edu

Abstract. This paper presents GasSense, a low-cost, single-point sensing solution for automatically identifying gas use down to its source (e.g., water heater, furnace, fireplace). This work adds a complementary sensing solution to the growing body of work in infrastructure-mediated sensing. GasSense analyzes the acoustic response of a home's government mandated gas regulator, which provides the unique capability of sensing both the individual appliance at which gas is currently being consumed as well as an estimate of the amount of gas flow. Our approach provides a number of appealing features including the ability to be easily and safely installed without the need of a professional. We deployed our solution in nine different homes and initial results show that GasSense has an average accuracy of 95.2% in identifying individual appliance usage.

Keywords: Ubiquitous Computing, Sustainability, Sensing, Gas.

1 Introduction and Motivation

Natural gas is the most widely consumed energy source in American homes [19]. It is used for furnaces, water heaters, stoves, fireplaces and, in some cases, clothes dryers. In the US, natural gas prices have quadrupled over the past decade due to growing demand and limited pipeline capacity [3]. As a result, government agencies and gas utilities have scrambled to implement conservation programs to reduce demand and better help customers manage energy costs (e.g., [7, 17]). Although recent work in the UbiComp and Pervasive research communities has focused on sensing electricity and water usage in the home [10, 12, 15, 16, 18], little attention has been directed towards sensing natural or propane gas usage. Unlike electricity and water usage, which are often the result of direct human actions such as watching TV, doing laundry, or taking a shower, gas usage is dominated by automated systems like the furnace or hot water heater. This disconnect between activity and consumption leads to a lack of consumer understanding about how gas is used in the home and, in particular, which appliances are most responsible for this usage [14]. Most people simply have no means of judging their household gas consumption other than a monthly bill, which, even then, does not provide itemized details about *what* accounts for this consumption.

In this paper, we introduce GasSense, a significant step towards eliminating this knowledge gap by providing highly granular information about gas usage in the home. GasSense is a low-cost, *single-point* sensing solution that uses changes in acoustic intensities of gas events to automatically identify gas use down to its source appliance (e.g., water heater, furnace, fireplace) and provide estimates of gas flow. GasSense automatically classifies gas use down to its source based *on flow volume* and *rate-of-change*. Each gas device in the home draws a unique amount of gas when activated and the flow rate-of-change is based on the type of appliance and, to some extent, its location in the home (i.e., the pipe pathway to the gas appliance). Because GasSense analyzes the *acoustic* signatures of gas events, it does not require direct contact with the gas itself to perform its calculations. This is in contrast to traditional gas sensing approaches, which only provide aggregate measurements of usage and require in-line contact with the gas to operate.

Previous work in sensing in the home has been motivated by one of two independent concerns: (1) human activity sensing for assistive care environments (e.g., elder care monitoring) [6, 13, 18] or (2) in enabling highly detailed eco-feedback applications to reduce wasteful consumption practices [2, 15, 16]. Although automatically identifying gas usage in the home may indeed provide insights into corresponding human activities (e.g., stove use indicates cooking, long hot water use indicates showering), a large majority of gas use stems from automated mechanical systems (e.g., from a home's furnace and water heater), which may or may not directly correspond to a human's current activity.

Thus, a primary focus of our work is in supporting the second concern, that is, enabling new types of eco-feedback [11] about gas usage in the home. Such feedback may come in the form of redesigned bills, home internet portals, or ambient home displays. The key here is in providing utilities and consumers with sensing data that is more than just an aggregate number but rather an itemized breakdown of gas usage down to its source. This detailed data should allow residents to make informed decisions about the costs and benefits of how they consume gas in their home (e.g., the temperature settings on their hot water heater, using hot water in the clothes washer). For example, in a review of the past 25 years of research into the effects of feedback on electricity consumption, Fischer found that feedback resulted in typical energy savings of between 5 and 12% [8]. It is likely that gas usage feedback will result in a similar decrease. Without detailed measurement, however, these sorts of applications are not possible. The focus of GasSense, then, is providing an easy to install sensing system capable of estimating disaggregated gas usage.

Given the small number of natural gas appliances in each home, it is tempting to consider a *distributed direct sensing* approach (e.g., [6, 16, 20]) for sensing gas usage (e.g., installing a flow sensor behind each appliance). There are three potential challenges with this approach: first, it requires constructing sensors that are flexible and robust enough to fit a variety of pre-existing gas appliance models in a non-invasive way; second, it inherently involves multiple sensors, which increases both the technical complexity (e.g., network communications) and the complexity of deployment; finally, natural gas is a highly combustible compound, so we were particularly interested in pursuing sensing approaches that were safe and did not require the help of a professional for installation. To address these challenges, we adopted an *infrastructure-mediated sensing* approach (e.g., [12, 18]), which leverages a home's existing physical infrastructure to sense and infer higher-level information about events in the home.

GasSense is the first single-point sensing solution capable of inferring both the fixture source and the flow level of natural gas. In this paper, we show how a standard gas meter’s government mandated regulator can be instrumented with a simple microphone-based sensor to gather acoustic signals of gas usage. We provide an overview of our gas fixture identification and flow estimation algorithms and the results of an evaluation in nine diverse homes. In particular, we show that we can accurately detect and identify gas events with 95.2% accuracy. Unlike previous work in infrastructure-mediated electricity and water sensing, we focus on both isolated appliance usage *as well as* overlapping usage (i.e., multiple gas appliances in use at the same time). Of the 496 gas events that we collected, 175 were recorded in isolation and 321 were recorded with other appliances on.

2 Related Work

There have been two prevailing approaches to sensing home resource consumption at the appliance or fixture level: direct sensing, and infrastructure-mediated sensing (indirect sensing). Direct sensing is sensing at or near the point of consumption. For example, in ViridiScope, Kim et al. [16] use a combination of magnetic, acoustic, and light sensors to sense the internal power state of appliances. Similarly, Arroyo et al. developed the WaterBot system [2], which senses and provides feedback about water usage behaviors at the water fixture itself by using an inline water flow sensor. Others have used computer vision [1], microphones placed within the living environment [6], and many simple sensors dispersed throughout the home [20]. Although direct sensing can provide highly granular data, wide scale deployment of such an approach is often impractical. The installation and maintenance of many direct sensors can be cost prohibitive, direct sensing can also create significant privacy concerns and stigmatization (especially with cameras or microphones), and introduce concerns over the aesthetics in the home, eventually leading to adoption problems [4, 13].

Recent work has therefore examined home resource sensing using just a handful of inexpensive sensors and often just a single sensor at strategic locations in a home’s existing infrastructure (e.g., a home’s water [10, 12, 15] or electrical [18] infrastructures). In contrast to direct sensing, infrastructure mediated sensing leverages the existing home infrastructure to mediate the transduction of events. A primary goal of these systems is to reduce economic, aesthetic, installation, and maintenance barriers to adoption by reducing the cost and complexity of deploying and maintaining the sensing infrastructure. The key enabler in applying this approach to gas sensing is that the home’s gas infrastructure and, in particular, its gas meter is standardized across households due to government regulation. Thus, the gas meter offers a particularly attractive installation point, which is safe, accessible, and also provides an indirect means to sense gas usage.

We are unaware of any commercial gas monitoring solutions that attempt to provide appliance-level data on gas usage. Smart gas meters, which can be remotely read by utility companies, typically only provide aggregate data. Small, inline gas sensors are also commercially available (e.g., [5]), but suffer from many of the same problems as the direct sensing approach (e.g., they require installation by a professional or direct access to the underlying pipe infrastructure). Industrial applications, especially

for manufacturing systems, have motivated the development of highly granular ultrasonic gas flow meters (e.g., [1]) but these systems can cost over \$1000 USD and therefore are not well suited for residential usage.

To the best of our knowledge, acoustic sensing has not been used for appliance-level identification and gas flow estimation as it is in GasSense. In leak detection, however, acoustics play a critical role. For example, several handheld devices, which are marketed towards utility companies, use high frequency response microphones to detect ultrasonic waves, which are characteristically emitted from small orifices (leak holes) in long haul gas pipes. These acoustic solutions only detect the presence of a leak and do not try to characterize the flow or any appliance activity. In our approach, the acoustic response produced at the gas regulator is at a much lower frequency because the chamber and orifice are significantly larger than that of most leaks.

3 Background and Theory of Operation

Gas is delivered to homes through a pressurized piping infrastructure. High-pressure transmission pipelines move gas from the production company's cleaning plants to gas distribution stations. Regulators and control valves control the high-pressure gas as it moves along the pipeline. At city gate stations, regulators reduce the pipeline gas pressure to distribution pressure. To provide a constant, measurable gas pressure, regulators control the gas pressure just before it enters the gas meter and into the home (Figure 1). In the case of propane, the gas is stored in an on-site tank, and enters the home only through a pressure regulator, as propane is typically unmetered.

Gas regulators are mandated by US national code (ANSI code B109.4-1998) to deliver safe pressure levels to the home's piping system. Given these government regulations, there is a reasonable level of expectation that the gas regulators are both consistent and present across homes. We leverage this fact in our sensing approach. The regulator consists of a diaphragm with a spring-loaded case that controls the



Fig. 1. (left) A typical US gas meter and pressure regulator. (middle) The location of our microphone-based sensor on the pressure regulator relief vent. (right) The GasSense prototype and sensor mount attached to the relief vent.

amount of gas flow (Figure 2). If the regulator senses high or low pressure changes, it adjusts accordingly to restrict or increase gas flow. As an added safety feature, relief (or breather) valves exist to vent gas harmlessly if a line becomes over-pressurized or the regulator malfunctions. This relief valve is connected to the diaphragm chamber and expels the gas through an external steel tube (the relief vent). When gas is being consumed in the home it is possible to hear the gas flowing through the regulator at the inlet valve, which typically sounds like a slight hissing noise. This sound is amplified by the diaphragm chamber, which acts as a resonant cavity.

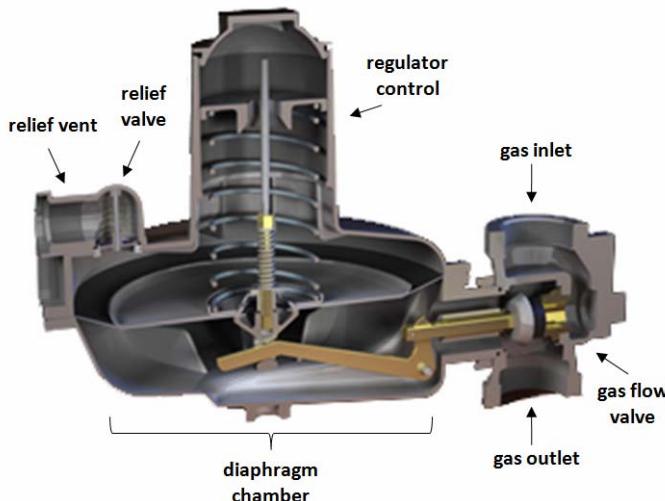


Fig. 2. A cross-section of a typical regulator design. We sense gas flowing through the gas flow valve via a microphone placed on the relief vent.

Although a microphone may seem like a rather indirect way of measuring gas flow, it has several theoretical foundations. First, for a fixed chamber, the resonant frequency is also fixed, determined entirely by the size of the chamber [9]. This is analogous to resonance in a whistle—even when you blow harder into a whistle, you do not change the pitch, just the intensity of the sound. Second, we know that greater flow through the tube can only result from greater pressure at one end of the tube. If the flow of gas inside the tube is laminar, the relationship between pressure exerted at the tube ending is *linearly* related to the flow through the tube. Moreover, this pressure, which is directly proportional to flow, manifests as the amplitude of the resonant frequency, and is ideally suited for measurement via a microphone. The relief vent (visible in Figure 1, 2 and 3) of the regulator is the only opening in the diaphragm chamber for which the sound propagates outward into the environment, making it ideally suited for sensor placement. With proper filtering and de-noising techniques, this signal can be isolated and calibrated to reflect aggregate gas flow, even in the presence of ambient noise (Figure 4 and 5).

4 Hardware Implementation and Data Collection Details

GasSense consists of an electret condenser microphone attached to a small printed circuit board (PCB), which contains the required amplification circuitry. The microphone is omni-directional, and has a sensitivity of -44 dB over the frequency range 100 Hz to 10 kHz. The PCB is attached to a sensor mount, which fits over the end of the relief vent on the gas regulator. The mounting bracket ensures that the microphone is level, centered directly under the regulator vent, and also protects the microphone from wind, rain and dust. Figure 1 (right image) and Figure 3 (right image) show the microphone mounted to a natural gas regulator. The microphone PCB requires external power which is provided by two AA batteries in a plastic enclosure.

The initial GasSense prototype is connected to the microphone input of a laptop's soundcard, although future designs could be entirely self-contained with a microcontroller and wireless transmitter. We experimentally confirmed in the lab using a signal generator that the automatic gain control (AGC) on the laptop's integrated soundcard was not attenuating or distorting the signal.

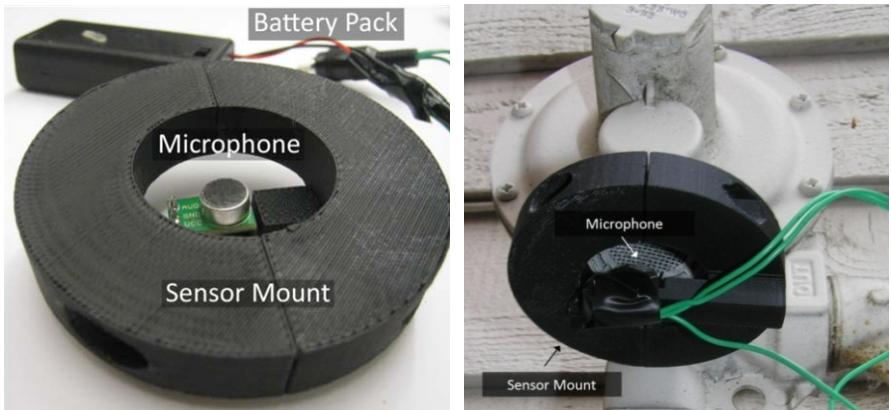


Fig. 3. (left) A close-up of our prototype system of GasSense with our sensor mount. (right) The GasSense unit attached to the relief vent of the pressure regulator.

4.1 In-Home Data Collection

In order to validate our sensing approach, we collected labeled data in nine different homes of varying size and age located in four cities (Table 1). Each house contained a varying level of background noise depending on the locality (proximity to a freeway, sidewalk traffic, etc.). In each house, we tested all of the available gas appliances including furnace, hot water heater, stove, fireplace (with both manual and electric starters), and a pool heater. Although our focus was on natural gas appliances, we also tested GasSense on one house that used propane (H5). For each dataset we noted the timestamp and appliance, which served as ground truth labels for evaluating our approach.

Table 1. Demographic data for homes used in our data collection experiments and their available appliances tested

House ID / Gas Type	Style / Built	Size / Floors	Furnace / Water Heater / Stove	Fireplace / Misc
H1 Natural Gas	Single-Family 1996	3400 sq. ft. 2	Yes / Yes / 4 burners	2 Manual Start
H2 Natural Gas	Single-Family 1998	3600 sq. ft. 2	Yes / Yes / 4 burners	2 Electric Start
H3 Natural Gas	Single-Family 1962	2000 sq. ft. 1	Yes / No / No	1 Manual Start / Pool Heater
H4 Natural Gas	Single-Family 2003	2900 sq. ft. 2	Yes / Yes / 6 burners	1 Electric Start
H5 Propane	Single-Family 1991	2100 sq. ft. 2	Yes / Yes / 4 burners	1 Electric Start
H6 Natural Gas	Single-Family ~1960s	1080 sq. ft. 1	Yes / Yes / 4 burners	No
H7 Natural Gas	Single-Family 1994	3360 sq. ft. 2	Yes / Yes / 4 burners	1 Manual Start
H8 Natural Gas	Single-Family 1991	3000 sq. ft. 2	Yes / Yes / 4 burners	1 Manual Start
H9 Natural Gas	Single-Family 1997	2600 sq. ft. 2	Yes / Yes / 4 burners	1 Electric Start

All of the data was collected and initially processed using the soundcard from our deployment laptops. The audio signal was sampled at 22,050 samples per second, enabling frequency analysis on the entire frequency range of the microphone. The raw data was recorded in an uncompressed WAV file using a 16-bit integer to represent each sample.

We followed a predefined experimental procedure to ensure that our data was consistent across deployment sites. For each home, we first attached the sensor to the relief vent on the gas regulator, as shown in Figure 1 and 3. We then individually turned each gas appliance on for a minimum of 15 seconds and then off. This was repeated at least three times for each appliance. This procedure allowed us to acquire a clean dataset where each appliance was the only gas device on in the entire house.

Many gas devices do not provide a mechanism to control the amount of gas flow—the device is simply on or off (furnaces, water heaters, dryers and some fireplaces). For all of the devices for which we could control flow, we slowly adjusted the device through each flow level to capture the effect of variable rate devices. For example, with gas stoves we would ramp a single burner from maximum to minimum flow and back again (Figure 4).

In addition to testing each appliance individually, we collected data involving more realistic scenarios in which more than one gas appliance was in use simultaneously (e.g., we would activate the furnace, water heater, and stove). These *compound events* are likely to be a common occurrence in any home and thus require special attention. To simulate compound events, we turned on multiple gas appliances at 15 second intervals up to four appliances at a time. Figure 4 (right image) shows data collected from a compound event test.

We also collected flow rate information for both automatically controlled (furnace and water heater) and manually controlled (stove and fireplace) gas appliances. We

used two methods for collecting ground truth gas flow: the natural gas meter and gas appliance labels/manuals, neither of which, unfortunately, provided perfectly accurate ground truth. Although natural gas meters do indeed provide measurements of gas flow, they are not designed to visualize accurate data about *instantaneous* flow. Even when a constant rate of gas is flowing, gas meter dials would commonly stutter and then later jump as much as a whole turn. To mitigate these effects, we collected flow measurement data over longer periods—typically four minutes or more, which corresponds to at least two cubic feet of gas—and averaged the results to obtain flow rate. This method was used on all homes except H5, which used unmetered propane.

As an alternative to the gas meter, we also used consumption information listed on the gas appliance (or its manual) for ground truth. Most large appliances are directly labeled with their power consumption (typically in BTU/hr). Using this method provides estimates for the gas consumption for individual appliances. However, the power consumption obtained from the appliance documentation cannot simply be converted into a gas flow rate as this conversion varies with temperature, pressure, and humidity. Therefore, this method of calibration can only give rough estimates of usage. For homes in which there is no meter (e.g. propane gas homes) this may be the only method to estimate the flow. For our analysis we used this method on H5, the only residence using unmetered propane.

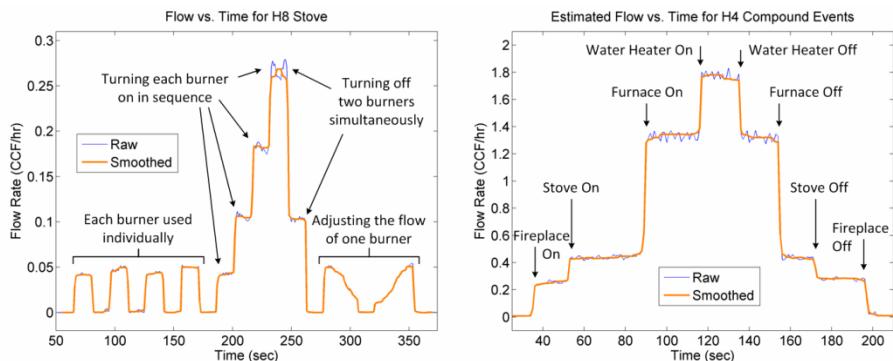


Fig. 4. Flow rate sensed from a calibrated microphone during a typical test deployment

5 Flow and Event Detection

We pursued a three-step approach to transforming the collected audio data into gas flow estimates and inferring appliance-level activity. First, the raw audio files are cleaned using Fourier transforms and band pass filtering. Second, the linear relationship between acoustic intensity and gas flow are used to estimate the amount of gas usage. Finally, the calculated gas flow *volume* and *rate-of-change* are used to classify the appliance source.

5.1 Initial Processing

The initial processing step involves preparing the raw audio files for analysis. For each home, the uncompressed WAV recording of the pressure regulator is grouped into non-overlapping time windows containing one second of data each. Using these one second windows, we compute a short-time Fourier transform (i.e., a spectrogram) of the audio signal. Figure 5 shows the resulting spectrogram, in decibels, as a function of both time and frequency. Notice that, as expected, the hissing sound primarily occupies a specific frequency in the spectrogram (in Figure 5, this is at 7.5 kHz). The frequency of resonance can change from regulator to regulator, but for all homes for which we have collected data, the resonance frequency stays in the range of 5 – 9 kHz.

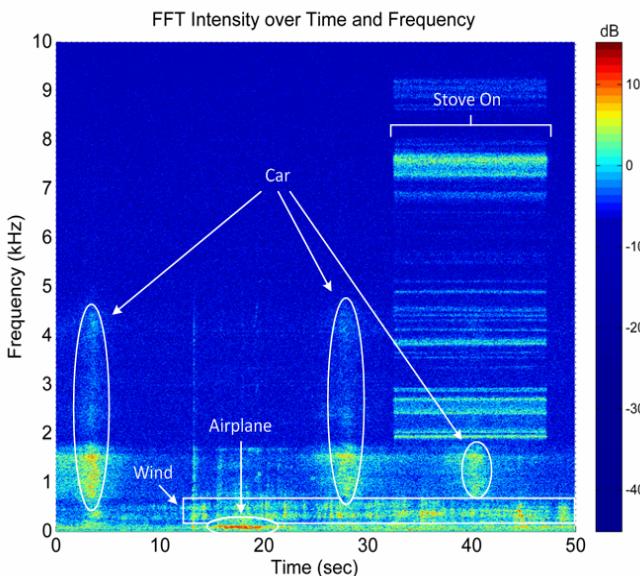


Fig. 5. Frequency spectrogram of ambient noise and regulator activity for a stove appliance from Home 9

Like all microphone-based approaches, our audio signal is susceptible to ambient noise. Fortunately, most of this noise (e.g., footsteps, wind, lawn-mowers) is low frequency (below 4 kHz)—note the wind, airplane and car noise in Figure 5. When noise *does penetrate* the 5 – 9 kHz range, it often has a wide-band frequency signature, making it easy to identify. In particular, if the mean energy outside the 5 – 9 kHz range is greater than one tenth of the energy at the resonance frequency, we discard the audio frame during this time and replace it with the median values from two seconds before and two seconds after the noise event (left image in Figure 6).

After removing environmental noise from the spectrogram, we find the resonant frequency by taking the maximum resonance in the range of 5 – 9 kHz. We extract the magnitudes of the resonant frequency across time to form a time-series vector of values (time, resonant frequency intensity) that correlates directly to flow (Figure 6). This vector is then smoothed using a moving average filter of five second length.

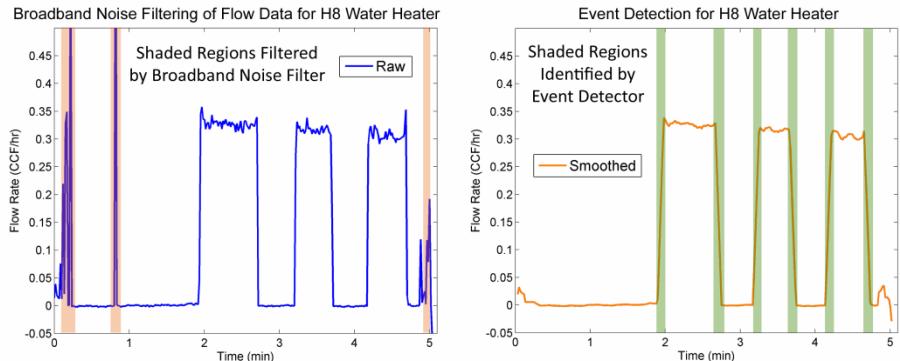


Fig. 6. (left) Raw audio signal with broadband noise detection. (right) Step change detector.

5.2 Gas Flow and Consumption

In order to use our acoustic measurements to estimate gas flow, we must first show that the acoustic signals correspond in a predictable way with the rate of consumption. As described in the theory of operation section, we expect the audio intensity of gas passing over the pressure regulator valve to be *linearly* related to flow. Figure 7 (left image) plots the audio intensity vs. ground truth measurements of gas flow for Home 4, which is representative of our dataset as a whole. Note that because of the aforementioned challenges with ground truth data collection, this graph is not perfectly linear, particularly for low-flow values, which are especially difficult to measure. As a result, we devised another test for linearity.

If there is a linear relationship between audio intensity and gas flow, the overall audio intensity of a compound event should simply be the sum of its individually collected audio intensity parts. This is in the same way that overall gas flow should be the sum of the individual flows from each appliance. To test this, we use our individually collected (single-event) dataset to compare with our compound event dataset. Figure 7 (right) plots the relationship between the expected value based on summing the individual appliance intensities with measured audio intensity for both individual and compound events. Note that the points lie on a unity slope indicating that relationship between audio intensity and rate of gas flow is in-fact linear. The slight non-linearity at high gas flow rates can be attributed to the microphone op-amp distorting very high amplitude signals (i.e., slew rate distortion), which can be compensated.

Using either the flow rate obtained from the meter or the consumption rate obtained from the appliance label/manual, the audio intensity can be calibrated to absolute units of gas flow (such as in CCF or Therms). Because the relationship between intensity and flow is linear, we use a simple linear regression to map intensity to flow, measured in Centum Cubic Feet (CCF or 100 cubic feet). The regression requires that we either have two points on the flow vs. intensity graph or only one point and assume that the origin (background noise level) is part of the dataset. Thus, we can calibrate the entire system from two appliances with different flow rates or from a single appliance with a variable flow rate. Of course, providing additional data points

from additional appliances can improve the regression and, consequently, the gas flow estimate.

At very high volumes (when all appliances are on) the microphone op-amp begins to distort the signal and introduce a non-linearity due to its insufficient dynamic range. Future deployments will include a gain control to automatically reduce the volume if distortion occurs. After data analysis, it was discovered that this non-linearity was significant in H8 and H9 where the sound produced by the regulator was louder than other homes. To compensate for this effect, a quadratic regression was used to map intensity to flow. The data used to form the regression equation was not used for testing.

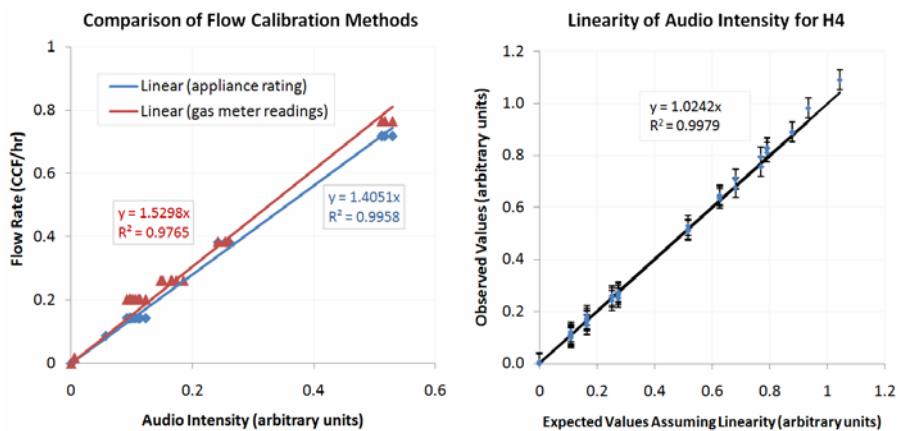


Fig. 7. (left) The linearity curve obtained from calibrating using the gas meter and appliance ratings as ground truth. (right) The linearity curve obtained from plotting observed step increases vs. expected increases in the audio signal.

5.3 Appliance-Level Event Detection and Identification

We use the smoothed resonant frequency vector as input to both our gas event detection and appliance identification algorithms. To identify gas usage events, we apply a sliding window step detector, which continuously looks for changes in magnitude of the resonant frequency intensity (e.g., Figure 4 or right image in Figure 6). The step detector triggers when it encounters a monotonically increasing or decreasing signal with a rate of change that is greater than a learned threshold. The threshold was determined for each house during calibration. The threshold is set to an arbitrarily large number and decreased in small steps. For each step, we segment a random subset of events that occur in isolation. If the correct number of events is calculated, the threshold is accepted. If not, the threshold is decreased and the process repeats. For example, if the subset contains four events, there should be four step increases and four step decreases segmented. We monotonically decrease the detection threshold until this pattern is seen. In this way, the threshold is set with minimal supervision.

After the step detector locates a step in the signal, the signal is passed to the appliance identification algorithm, which extracts three features: (1) the relative magnitude

of the step change in flow, (2) the slope of the change-in-flow, (3) and the rise or fall time of the unknown event. The first feature (the step size) provides an estimate of the amount of gas the appliance consumes. This feature is useful in disambiguating appliances that have fixed flow rates (e.g., a water heater typically uses less gas than a furnace). Interestingly, the step size is also useful in disambiguating appliances that have variable flow (e.g., a stove or fireplace). This is because these systems are designed to turn on at maximum flow during activation, providing a reliable step increase. The second and third features (the step slope and rise time) are useful because electromechanically switched appliances have very steep operating slopes when compared with manually or human controlled appliances. Though it may seem that the slope does not add any additional information given the step change and rise time as features, we found that some of our datasets performed better by as much as 4.6% when using the slope.

Feature vectors are generated for each segmented event and then used to build a k-nearest neighbor (KNN) model using the Weka Toolkit [21]. KNN is used to automatically determine the source of the gas events. We applied KNN ($k = 3$) with a Euclidean distance metric and inverse weighting, which is well suited for this kind of feature vector because a small distance in the N-dimensional space corresponds to gas events having similar flow and slope. These parameters were derived through experimentation and cross validation of our entire dataset.

6 Performance Evaluation and Results

Our total dataset includes 496 gas events collected over nine homes and five separate appliance types (furnace, water heater, stove, fireplace, and pool heater). Of the 496 gas events, 175 were recorded in isolation and 321 were recorded in compound. In this section, we present the results of our event detection and appliance identification algorithms.

6.1 Event Detection

To evaluate the accuracy of our event detection algorithm, we iterated over each gas appliance event in our dataset (including both single and compound events). We compared the output of the event detection algorithm to our ground truth labels. We were able to correctly detect 98.22% (1st column in Table 2) of all gas events, even in the presence of considerable ambient noise. For example, leaf blowers, passing cars, and speech were present in many of our datasets, but were not the cause of any failures. In fact, all homes, with the exception of H5, had 100% accuracy. H5, however, is a special case—it is the only home in our dataset that used propane rather than utility supplied natural gas. Unlike natural gas meters, which rely on a single pressure regulator to stabilize incoming gas, propane homes use two regulators (for safety and efficiency reasons). Since these two regulators do not regulate the pressure proportionally, the quantity of gas flow cannot accurately be determined by monitoring only one regulator. The gas stove events, for example, were completely missed on four separate trials.

6.2 Appliance Classification

Once a gas event is detected and isolated, its features are sent to our appliance classification algorithm to identify the source of the event. To test the accuracy of our appliance classification algorithm, we ran a 10-fold cross validation experiment across each detected gas event for every home. The 2nd column in Table 2 shows the results of this experiment. The aggregate accuracy across all homes was 95.2% and the worst performing home, H5, was still above 85% accuracy. Again, the issue with H5 was with the intensity of the audio signal when monitoring only one of the two propane gas regulators.

Table 2. Overall performance of event detection and classification categorized by home

Home (N=# of gas events collected)	Events Detected	10-Fold Cross Validation Classification Results	Classification Results Using Minimal Training Set
H1 (N=72)	100%	93.05%	88.24%
H2 (N=87)	100%	98.85%	99.05%
H3 (N=24)	100%	100%	100%
H4 (N=102)	100%	95.07%	98.34%
H5 (N=50)	84%	85.71%	86.12%
H6 (N=22)	100%	100%	83.34%
H7 (N=32)	100%	100%	100%
H8 (N=58)	100%	87.75%	80.98%
H9 (N=49)	100%	96.55%	49.8%
Aggregate (N=496)	98.22%	95.22%	87.32%

Table 3. Overall performance of event detection and classification categorized by fixture

Appliance Type (N=# of gas events collected)	10-Fold Cross Validation Classification Results	Classification Results Using Minimal Training Set	
		Appliance On	Appliance Off
Furnace (N=108)	98.13%	100%	98.5%
Water Heater (N=88)	93.02%	100%	76.1%
Stove (N=206)	97.56%	96.8%	84.8%
Fireplace (N=88)	91.66%	100%	93.2%
Pool Heater (N=6)	100%	100%	100%
Aggregate (N=496)	96.07%	99.36%	90.52%

Table 4. Confusion matrix from 10-fold cross validation classification

	Fireplace	Furnace	Stove	Water Heater	Pool Heater
Fireplace	77	0	2	5	0
Furnace	1	105	0	1	0
Stove	3	0	200	2	0
Water Heater	4	2	0	80	0
Pool Heater	0	0	0	0	6

Table 3 presents an alternate view of the same data as in Table 2 but categorized by appliance rather than by home. The cross validation reveals that the poorest appliance classification accuracies involved fireplaces and water heaters. Upon further analysis we observed that 75% of the incorrect water heater classifications, were misclassified as fireplace events. As we discuss later, there is a characteristic low frequency thump, which can be used to differentiate fireplace and water heater events. Note that our data was labeled according to individual appliance, not appliance type. This enabled us to investigate whether we could automatically distinguish between appliances of the same type but in different locations in the home. The H2 dataset, for example, contained *two different* fireplaces that were correctly distinguished from one another with 100% accuracy.

The 10-fold cross validation shows that our particular KNN-based classifier performed well at correctly classifying gas events down to the source appliance. Table 4 presents the confusion matrix summarizing the classifier's prediction. However, in real-world deployment scenarios, our training dataset will likely be smaller. That is, a homeowner would likely only be willing to provide one example use of each appliance. To test this sort of scenario, we trained solely on one or two individual gas events for each appliance and tested on the rest of the dataset. These results are presented in the far right column of Table 2 and the last two columns in Table 3. The low accuracy in H9 is the result of the introduction of non-linearity to the sensed signal, which affected both H8 and H9 (as mentioned in Section 5.2). The induced non-linearity in the dataset made it impossible to find a calibration subset that effectively represented all data collected. As a result, step increases during compound events were not representative of the trained step increases. After thorough analysis, we found that this was due to the low dynamic range of the microphone op-amp. Future deployments will include a gain control to automatically reduce the volume if distortion occurs. We also noticed that as more gas appliances are running simultaneously (i.e., higher overall flow), the noise is also amplified, which explains some of the misclassifications for compounded events (especially small loads). Unlike the minimal training set results, H9 performs well in cross validation. This can be explained by the larger set of training data. Cross validation randomly selects 90% of total instances for training and rest for testing in each fold. This results in a classifier that learns from the nonlinear instances too, resulting in a more robust model.

7 Discussion

Our initial results show significant promise for appliance-level single-point sensing of all gas appliances in the home using a simple microphone-based sensing approach. After taking data from nine homes, we have shown that individual appliances can be reliably detected and classified with an average overall accuracy of 95.2%. This is an important advancement in gas monitoring, and we are not aware of any prior work in appliance-level single-point sensing of gas. Our microphone-based approach has the added advantage of being safely installed by a non-professional. In this section, we discuss potential implementations for the end-user calibration process, alternative sensing approaches we explored, and outline the next steps and potential future work.

7.1 End-User Calibration

Our approach requires two training procedures by an end user: the association of relative flow produced by GasSense to absolute flow volume and providing ground truth labels of appliances to their respective audio signatures. We can imagine users employing one of three calibration methods for associating the relative flow inferred by our sensing approach to absolute consumption (such as in CCF or Therms). The first would involve using the readings from the home's gas meter as we did in our own experiments. The second involves reading the flow ratings on certain appliances, such as the water heater or furnace (again, this is an approach we used in our experiments). Large appliances are typically required by national code to show their gas consumption. This method is particularly useful for homes that use propane as they do not typically have a gas meter. The third, less intensive method is to use the measurements reported on the gas bill. Since GasSense can record the duration of gas usage and its relative flow, we can use this term to calculate the total gas consumed over a period of time. In this way, the first gas bill (or even a set of sparse measurements from the meter) can be used to calibrate the system to absolute units of flow. The homeowner would only need to enter the dates of use and aggregate gas usage. Although many gas utilities in the United States charge gas consumption in units of energy (typically in Therms), the bill also reports the total volume of gas measured by the meter (typically in CCF), enabling the use of this type calibration method. For associating appliance labels to events, we can imagine the user simply activating each appliance in sequence and then entering the sequence of labels through a user interface.

7.2 Alternative Sensing Solutions

Before deciding on a microphone-based sensing solution, we explored the idea of using a methane sensor placed in the relief vent. This solution proved to be inadequate due to the sensitivity and response time of the sensor. The methane sensing approach relies on trace amounts of gas that escape from the regulator during *changes* in gas flow. However, for many appliances the flow is not great enough to exert backpressure on the regulator, and no methane escapes from the relief valve. Moreover, pellistor and infrared methane gas sensors both suffer from inadequate response times. Because we use the slope of the step increase in flow to classify appliances, response time proved to be a crucial issue. We also experimented with a thermistor placed just inside the relief vent. The idea behind this approach is that as the escaping gas leaves the pressurized atmosphere of the regulator, it cools as it expands into the ambient air. Unfortunately, the volume of gas released through the relief valve was not sufficient to reliably affect the thermistor.

7.3 Additional Potential Features

During our experimentation we also observed a potentially useful feature of furnaces and hot water heaters that could be included into future classification. Furnaces and water heaters produce a low frequency thump from their ignition modules when the valve opens and closes (Figure 8). These appliances have two-state solenoid valves to mechanically control the flow of gas, producing a characteristic thump as the solenoid slams the valve into position. This thump is easily sensed using the microphone, and

varies significantly between appliances which may be used to help differentiate similar events. Unfortunately, the microphone also picks up substantial low frequency noise from people walking outside the home, cars passing by or other sources from the ambient environment. This makes feature extraction at these frequency levels inherently unreliable. Noise cancelling hardware would likely help alleviate this problem. Instead of probing the lower frequencies for features continuously, we envision using the low frequency analysis to disambiguate appliances that cannot reliably be distinguished based solely on flow. In H1, for instance, this feature would help to disambiguate the water heater from the fireplace, which were sometimes confused for one another.

A second interesting structural pattern that we observed during our analysis of the signal was a low frequency modulation of the intensity of the hissing sound. At first we speculated that it could be an appliance specific phenomenon, but after conducting a thorough analysis, we attributed it to the mechanical operation of the meter. Household gas meters are positive displacement diaphragm meters where the gas flows through two chambers making a piston move. The gas needs to exert pressure on these pistons for this movement, which in turn creates a small backpressure on the inlet side causing the regulator's diaphragm to slightly move to compensate. As the gas flows from one chamber to another, the pistons move back to their original position. This repeated motion manifests itself as an oscillating pressure change.

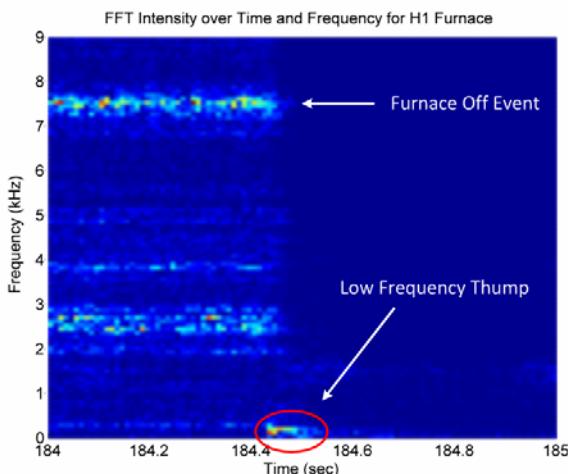


Fig. 8. An example of a low frequency thump caused by the ignition module in a gas furnace

7.4 Addressing Limitations and Future Work

GasSense has been primarily tested with natural gas meters supplied by the public utility; however, a single home was tested using a propane tank, which demonstrates that GasSense may also be a viable option for propane. Most homes with propane have two regulators instead of one, a regulator at the outlet of the tank and one at the inlet to the house. The additional regulator causes several complications for GasSense

because it cannot be assumed that the task of regulating the pressure is shared equally by the two regulators, and it therefore may be necessary to monitor both regulators.

In future implementations, we will experiment with using a directional microphone to eliminate some background noise. In addition, non-linear signal distortion will be remedied with a higher dynamic range microphone op-amp and a gain control to automatically reduce the volume before distortion occurs. We will explore the use a second microphone near the main microphone, but not underneath the vent of the regulator, which can be used for noise cancellation. This would dramatically increase our signal to noise ratio and allow us to measure extremely quiet hissing.

We also plan to perform long-term deployments of GasSense to explore the reliability of the system over time in a naturalistic usage setting. In particular, we would like to study the effects of the environmental variables such as temperature, humidity, and barometric pressure on the acoustic signal produced by the regulator. Additionally, we would like to explore the possibility of classifying events that occur at the exact same time, despite the low probability of this occurrence (e.g., a furnace and hot water heater turning on simultaneously).

8 Conclusion

In this paper we have presented GasSense, a new complementary infrastructure-mediated sensing solution for the gas infrastructure. GasSense is extremely cost-effective using only a commodity microphone for its sensing approach and provides a novel single-point solution for sensing gas use down to the appliance level. Our approach provides a number of appealing features including having the ability to be easily and safely installed without the need of a professional. We deployed our solution in nine different homes and found that a KNN classifier could be used to classify audio signals to their appliance source with an accuracy of 95.2%. We hope to combine this solution with past infrastructure-mediated sensing systems to provide a complete picture of whole-house activity as well as support new eco-feedback applications that provide users with disaggregated energy consumption information.

References

1. Ao, X., Matson, J., Kucmas, P., Khrakovsky, O., Li, X.: UltraSonic Clamp-On Flow Measurement of Natural Gas, Steam, and Compressed Air, <http://www.gesensing.com/products/resources/whitepapers/ur268.pdf> (last accessed 10/16/2009)
2. Arroyo, E., Bonanni, L., Selker, T.: Waterbot: exploring feedback and persuasive techniques at the sink. In: CHI 2005, pp. 631–639. ACM, New York (2005)
3. Balasch, P.: National Gas and Electricity Costs and Impacts on Industry. National Energy Technology Laboratory, prepared for US Dept. of Energy, DOE/NETL-2008/1320 (2008)
4. Beckmann, C., Consolvo, S., LaMarca, A.: Some Assembly Required: Supporting End-User Sensor Installation in Domestic Ubiquitous Computing Environments. In: Davies, N., Mynatt, E.D., Siio, I. (eds.) UbiComp 2004. LNCS, vol. 3205, pp. 107–124. Springer, Heidelberg (2004)

5. Captor In-line Type Flow Meter,
<http://www.captor.com/> (last accessed 10/16/2009)
6. Chen, J., Kam, A.H., Zhang, J., Liu, N., Shue, L.: Bathroom Activity Monitoring Based on Sound. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) PERVASIVE 2005. LNCS, vol. 3468, pp. 47–61. Springer, Heidelberg (2005)
7. Chicagoland Natural Gas Savings Program,
http://www.conservationsrebates.com/programs/chi/CHI_Index.aspx (last accessed on 10/16/2009)
8. Fischer, C.: Feedback on household electricity consumption: a tool for saving energy? Energy Efficiency 1, 79–104 (2008)
9. Flanagan, J.: Speech analysis synthesis and perception. Springer, New York (1972)
10. Fogarty, J., Au, C., Hudson, S.E.: Sensing from the Basement: A Feasibility Study of Unobtrusive and Low-Cost Home Activity Recognition. In: UIST 2006, pp. 91–100 (2006)
11. Froehlich, J., Findlater, L., Landay, J.: The Design of Eco-Feedback Technology. In: CHI 2010 (to appear 2010)
12. Froehlich, J., Larson, E., Campbell, T., Haggerty, C., Fogarty, J., Patel, S.N.: HydroSense: infrastructure-mediated single-point sensing of whole-home water activity. In: UbiComp 2009, pp. 235–244 (2009)
13. Hirsch, T., Forlizzi, J., Hyder, E., Goetz, J., Kurtz, C., Stroback, J.: The ELDer Project: Social and Emotional Factors in the Design of Eldercare Technologies. In: Conference on Universal Usability, CUU 2000, pp. 29–72. ACM, New York (2000)
14. Kempton, W., Layne, L.: The Consumer's Energy Analysis Environment. Energy Policy 22(10), 857–866 (1994)
15. Kim, Y., Schmid, T., Charbiwala, Z.M., Friedman, J., Srivastava, M.B.: NAWMS: Non-Intrusive Autonomous Water Monitoring System. In: Conference on Embedded Network Sensor Systems, SenSys 2008, pp. 309–322. ACM, New York (2008)
16. Kim, Y., Schmid, T., Charbiwala, Z., Srivastava, M.B.: ViridiScope: design and implementation of a fine grained power monitoring system for homes. In: UbiComp 2009, pp. 245–254 (2009)
17. Natural Gas Conservation and Ratemaking Efficiency Act § 56-600 et seq. Virginia (2009)
18. Patel, S.N., Robertson, T., Kientz, J.A., Reynolds, M.S., Abowd, G.D.: At the Flick of a Switch: Detecting and Classifying Unique Electrical Events on the Residential Power Line. In: Krumm, J., Abowd, G.D., Seneviratne, A., Strang, T. (eds.) UbiComp 2007. LNCS, vol. 4717, pp. 271–288. Springer, Heidelberg (2007)
19. US Energy Information Administration, Using and Saving Energy in Homes,
http://tonto.eia.doe.gov/kids/energy.cfm?page=us_energy_homes (last accessed 10/16/2009)
20. Wilson, D., Atkeson, C.G.: STAR: Simultaneous Tracking & Activity Recognition Using Many Anonymous Binary Sensors. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) PERVASIVE 2005. LNCS, vol. 3468, pp. 62–79. Springer, Heidelberg (2005)
21. Witten, I., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)

Transferring Knowledge of Activity Recognition across Sensor Networks

T.L.M. van Kasteren, G. Englebienne, and B.J.A. Kröse

Intelligent Systems Lab Amsterdam
University of Amsterdam
Science Park 107,1098 XG, Amsterdam
The Netherlands
T.L.M.vanKasteren@uva.nl
<http://www.science.uva.nl/~tlmkaste>

Abstract. A problem in performing activity recognition on a large scale (i.e. in many homes) is that a labelled data set needs to be recorded for each house activity recognition is performed in. This is because most models for activity recognition require labelled data to learn their parameters. In this paper we introduce a transfer learning method for activity recognition which allows the use of existing labelled data sets of various homes to learn the parameters of a model applied in a new home. We evaluate our method using three large real world data sets and show our approach achieves good classification performance in a home for which little or no labelled data is available.

1 Introduction

Automatically recognizing activities, such as cooking, sleeping and bathing, in a home setting allows many applications in areas such as intelligent environments [27] and healthcare [122,29]. It is to be foreseen that in the near future activity recognition systems will be installed on a large scale (i.e. in many homes). Most state of the art activity recognition models are supervised models that require labelled data to learn the model parameters [10,16,25,27]. Because of differences in both the layout of houses and the behaviour of their inhabitants, a model trained for one house cannot be used for another house. This means that a labelled dataset needs to be recorded for each house. Since this is expensive, we propose to use transfer learning [35,24] to transfer knowledge from labelled datasets to situations where no or little labelled data is available.

Transfer learning has been successfully applied to independent and identically distributed (i.i.d.) data, using discriminative models [14,21]. Activity recognition, however, presents us with two important challenges: First, our measurements are part of a time series, and are therefore not i.i.d. Second, we deal with situations where the data from a house is largely unlabelled, hence making discriminative models inadequate. In this paper, we propose a method for applying transfer learning to time series (where the data points are not independent), using a generative model to allow the use of both labelled and unlabelled data

during learning. We apply our method in the health care domain where the goal is to recognize activities of daily living (ADL) from wireless sensor network data. The list of ADLs is a well recognized fixed list of activities which are good indicators for the cognitive and physical wellbeing of elderly [13]. In our experiments we recognize the same set of ADLs in different houses, having different sensor networks. Three large real world datasets are used to evaluate the performance of our method in activity recognition.

The rest of this paper is organized as follows. Chapter 2 describes related work of both activity recognition and transfer learning. In chapter 3 we describe our transfer learning approach in detail. Chapter 4 discusses the experiments and results. Finally, in chapter 5 we sum up our conclusions.

2 Related Work

This section describes the related work of activity recognition systems and transfer learning approaches. Furthermore, the terminology is introduced which is used throughout the rest of the paper.

2.1 Activity Recognition Systems

Activity recognition systems consist of a sensing system for obtaining observations and a recognition model which interprets these observations and recognizes which activities are performed. Sensing systems may include camera's [10], RFID [19][30], wearables [11][15] and wireless sensor networks [23][27].

Several models for activity recognition have been proposed, mainly of probabilistic nature. Good results are obtained using generative models such as the hidden Markov model (HMM) [19][27] and discriminative models such as conditional random fields (CRF) [6][27]. Extensions to these models such as hierarchical models [16][18] and segment models [10][25], have been proposed to deal with the long term dependencies in activities.

All these models are supervised models and therefore require labelled data to learn the model parameters. Some models have been proposed that somewhat reduce the need for supervised data such as a hybrid generative and discriminative model [12] or models that use common-sense knowledge from the web [31]. Such models provide interesting new opportunities for modelling activity recognition. However, the advantage of our method is that any existing or upcoming generative model that has proven itself in the field of activity recognition can be used without altering the model. That is, we can simply use the proposed model and learn its parameters using our transfer learning approach.

2.2 Transfer Learning

Transfer learning refers to techniques that learn model parameters for a classification task by incorporating training data from different, but related classification tasks. We distinguish between *source* tasks that provide us with training data,

and a *target* task which is the actual classification task we are interested in. Early work on transfer learning primarily focused on multi-task learning in which several tasks were learned jointly, yielding a better performance than learning the tasks separately [3,5,24].

For example, the goal in newsgroup classification tasks is to classify which newsgroup a particular document belongs [9,21]. One task is to recognize if a document comes from a newsgroup about space or about hardware. When including training data of other newsgroups such as religion, baseball and motorcycles the performance improves significantly [21]. This is because the other newsgroups provide information about the co-occurrence of words. A word such as ‘moon’ might often occur together with the word ‘rocket’. If the word ‘rocket’ did not occur in the space newsgroup dataset but the word ‘moon’ did, the classifier can still learn that ‘rocket’ is descriptive for the space newsgroup. Because it occurs often together with ‘moon’ in the other datasets.

The optimal way to perform transfer learning is still an active topic of research. One approach that seems to work well with probabilistic models is the use of a prior distribution over the model parameters. The prior provides an initial estimate of the model parameters for target task and is learned from the source tasks [14,21]. The influence of the prior decreases as more training data is observed, therefore providing a natural mechanism to balance the effect of the prior distribution and the training data while learning the model parameters.

3 Transfer Learning for Activity Recognition

When applying transfer learning to activity recognition each classification task corresponds to a house in which we perform activity recognition. We distinguish between a target house, for which there is little or no training data available, and a number of source houses for which we have large labelled datasets. The same list of ADLs is used for each house, while the sensor network for each house is different. To perform transfer learning from the source houses to the target house, two problems need to be solved:

1. How do we deal with differences in sensor networks due to the different layout of houses?
2. How can we learn model parameters such that differences in behaviour of the inhabitants are taken into account?

The first problem involves differences in feature spaces between the houses. Because each house has a different layout, the sensor network in each house has a different configuration, resulting in a different feature space. For example, one house might have a separate room for the toilet and bathroom, while these may be built together in another. As a result, the sensors used will differ, both in number and in function. To solve this we need to introduce some kind of mapping allowing us to have a single common feature space that can be used for all houses. We use meta features [14] for this mapping, which are features that describe the properties of the actual features. Each sensor is described by one

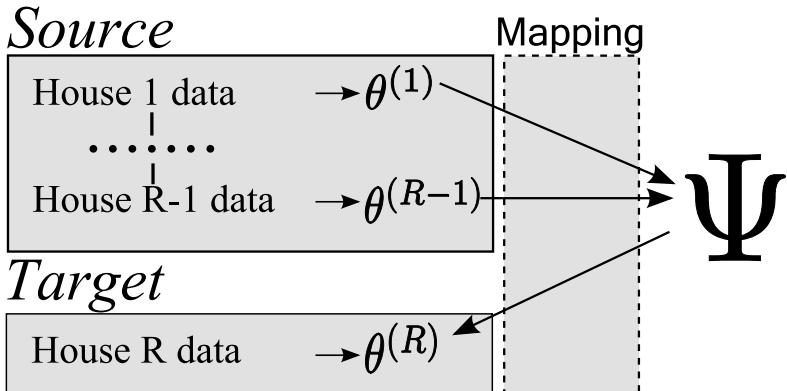


Fig. 1. Graphical representation of transfer learning framework. For each source house i training data is used to learn model parameters $\theta^{(i)}$. All the source model parameters are used to learn the hyperparameters Ψ of the prior distributions. Which in turn is used to learn the target model parameters $\theta^{(R)}$ together with any available data from that house.

or more meta features, for example, a sensor on the microwave might have one meta feature describing the sensor is located in the kitchen, and another that the sensor is attached to a heating device.

The second problem involves differences in behaviour between inhabitants. Even though for each house the same activity labels are used, there may still be differences in how activities are performed. For example, one person might often have cereal for breakfast, while another prefers toast. Such differences in behaviour require different sets of parameters to allow the model to recognize the corresponding activities. Therefore, we use a separate model for each house, each having its own set of model parameters. A prior distribution is learned from the source houses and used to provide a sensible initial value for the model parameters of the target house. Specific behaviour can then be accounted for by further updating the parameters using unlabelled and/or labelled training data from the target house. The entire approach is shown in a diagram in Figure 1.

In the rest of this section we first explain the type of mapping and the activity recognition model that we use. Then we explain how we learn the prior distribution and how this prior distribution is used to learn the parameters of the target model.

3.1 Mapping Using Meta Features

We define a sensor feature space as the original feature space in which each sensor of a house represents a feature, while the meta feature space is represented by meta features. There is a separate sensor feature space for each house, while the meta feature space shared by all houses. Choosing a proper mapping is difficult, since the optimal choice is not clear and a wrong decision can strongly

Table 1. Example of sensors (horizontal) being represented by meta features (vertical) for two houses

House	Sensors	Bathroom Entrance	Bathroom Other	Kitchen Heating	Kitchen Storage	Kitchen Other	Outside Entrance	Bedroom Entrance	Bedroom Other	Toilet
House A	Microwave	0	0	1	0	0	0	0	0	0
	Stove	0	0	1	0	0	0	0	0	0
	...									
House B	Microwave	0	0	1	0	0	0	0	0	0
	Refrigerator	0	0	0	1	0	0	0	0	0
	...									

affect the performance of the model. In previous work on transfer learning for activity recognition a comparison of mappings was made [26]. The mapping that combined sensor readings in a single feature based on their function (e.g. sensors used during cooking) gave the best results. We use the same type of mapping in the form of meta features, by defining meta features that describe the function of sensors (Table 1).

It is important to notice is that we do *not* first map all the sensor data from the sensor feature space to the meta feature space. Instead, as can be seen in Figure 1, this mapping occurs when learning the prior distribution, and using the prior to learn the target model parameters. However, it is possible to first map all the sensor data to the meta feature space and then create a single model for all houses. This approach was taken in [26] and we compare the performance of our approach to that approach in the experiment section.

3.2 Model for Activity Recognition

To recognize the activities from sensor data we use the hidden Markov model (HMM), which has been shown to perform well in this domain [19,27]. The HMM is defined in terms of an observable variable \mathbf{x}_t (the features in the sensor feature space) and a hidden variable y_t (the activities to recognize) at each time slice. Two dependency assumptions specify the model,

- The hidden variable at time t , namely y_t , depends only on the previous hidden variable y_{t-1} (*Markov assumption* [20]).
- The observable variable at time t , namely \mathbf{x}_t , depends only on the hidden variable y_t at that time slice.

These assumptions allow us to factorize the joint probability over all variables as follows

$$p(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}) = p(y_1) \prod_{t=1}^T p(\mathbf{x}_t | y_t) \prod_{t=2}^T p(y_t | y_{t-1}). \quad (1)$$

Table 2. Overview of the distributions used in the HMM, parameterized by the model parameters $\theta = \{\pi, A, B\}$. And the corresponding prior distribution, parameterized by the hyperparameters $\Psi = \{\eta, \rho, \omega, v\}$.

Factor	Model Distribution		Prior Distribution	
Name	Name	Parameters	Name	Hyperparameters
Initial State	Multinomial	π	Dirichlet	η
Transition	Multinomial	A	Dirichlet	ρ
Observation	Binomial	B	Beta	ω, v

The different factors represent: the initial state distribution $p(y_1)$ parameterized by π ; the observation distribution $p(x_t | y_t)$ parameterized by B ; the transition distribution $p(y_t | y_{t-1})$ parameterized by A . The entire model is therefore parameterized by a set of three parameters $\theta = \{\pi, A, B\}$.

For more technical details about distributions used in the HMM we refer the reader to appendix A.

3.3 Learning the Prior Distribution

In Bayesian statistics a prior is said to be conjugate if the resulting posterior is of the same functional form as the prior [4]. The parameters of prior are typically called hyperparameters Ψ , to clearly distinguish them from the model parameters θ . We use conjugate priors in this work, an overview of all the distributions and their parameters can be found in Table 2.

To learn the hyperparameters we first learn the model parameters of the source houses. Because we have large labelled datasets for the source houses, we can easily learn those parameters using maximum likelihood. These source model parameters provide us with examples of what the model parameters look like and are used to learn the hyperparameters.

Learning the hyperparameters of the initial state distribution and the transition distributions is straightforward, because the dimensionality of the model parameters is the same for all houses. We can calculate them efficiently using numerical methods [17].

Estimating the hyperparameters of the observation distributions is more involved because of the different sensor feature spaces in each house. This is where the earlier proposed mapping comes into play. We map the learned observation model parameters of the source houses to the meta feature space. Then we use those values to learn the hyperparameters using numerical methods.

For more technical details about learning the hyperparameters we refer the reader to appendix B.

3.4 Using the Prior to Learn the Target Model Parameters

To learn the model parameters of the target house we use the EM algorithm. In the E-step any available unlabelled and/or labelled data from the target house is used to calculate the expectations. During the M-step these expectations are

used to calculate the new set of parameters, only this time the prior distribution is added to that calculation.

For the initial state distribution and the transition distribution this is straightforward, because the hyperparameters are of the same dimensionality as the model parameters. However, the observation hyperparameters are stored in the meta feature space and therefore need to be mapped to the sensor feature space of the target house. The EM algorithm is run until it converges, after which transfer learning the target model parameters is completed.

For more technical details about learning the hyperparameters we refer the reader to appendix C

4 Experiments

We want to find a good method for transfer learning in activity recognition. Our proposed approach is characterized by three elements: 1. Meta features are used to map between the sensor feature space and to a common meta feature space; 2. A separate model is used for each house to take into account the differences in behaviour of the inhabitants and 3. A generative model is used to allow the inclusion of unlabelled data of the target house during the learning process.

To validate how well this approach works we perform the following experiments. First, we compare the performance of a model using the meta-feature space for representing the sensor data, to a model using the original sensor feature representation. Second, we compare the performance of using a separate model for each house, to the performance of a single model for all houses. Third, we compare the performance of using both labelled and unlabelled data, to the performance of using only labelled data.

We first give a description of the houses and the datasets recorded in them and provide details of our experimental setup. Then we present the results and discuss the outcome.

4.1 Sensor System

In this work we apply our method to wireless sensor network data, using both existing and novel real world datasets. Our wireless sensor network consists of several wireless network nodes to which sensors can be attached. Examples of sensors used include reed switches to measure open-close states of doors and cupboards; pressure mats to measure sitting on a couch or lying in bed; mercury contacts for movement of objects (e.g. drawers); passive infrared (PIR) to detect motion in a specific area; float sensors to measure the toilet being flushed. Sensor output is binary and represented in a feature space which is used by the model to recognize the activities performed.

4.2 Data

Our sensor system was used to record datasets in three houses. One three room apartment (house A), one two room apartment (house B) and one two story

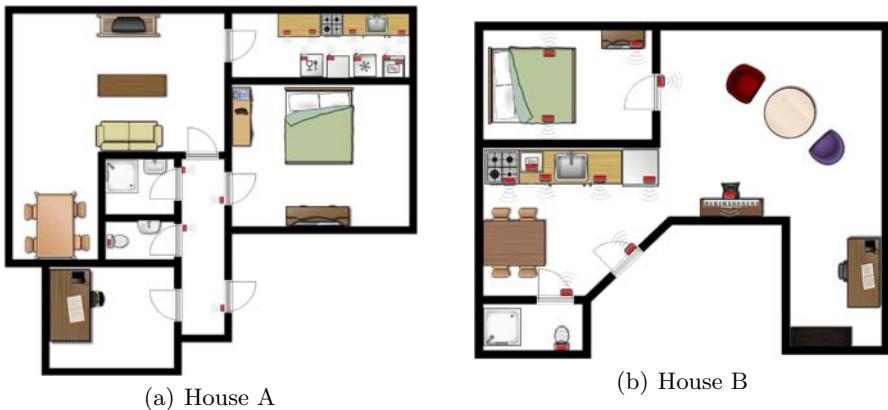


Fig. 2. Floorplan of houses A and B, the red boxes represent wireless sensor nodes.
(created using: <http://www.floorplanner.com/>)

house (house C), an overview of the datasets can be found in table 3. The datasets are available at: <http://www.science.uva.nl/~tلمکاسته>.

The layout of the houses differs strongly, for example, there are two toilets in house C, the toilet in house B is in the same room as the shower, while the toilet and shower in house A are in separate rooms. Furthermore, the inhabitants differ as well, house A was occupied by a 26 year old male, house B by a 28 year old male and house C by a 57 year old male. To further illustrate the differences between the houses we have included the floorplans of houses A and B (Fig. 2) and house C (Fig. 3).

We asked the inhabitants to annotate their behaviour using eight activities based on the list of activities of daily living (ADLs), a health care standard for monitoring elderly [13]. The activities in house A and B were annotated using a wireless bluetooth headset, the inhabitant recorded the start and end point of an activity while performing it. In house C activities were annotated using a handwritten diary. The activities annotated are the same for all three houses

Table 3. Information about the datasets recorded in three different homes using a wireless sensor network

House	House A	House B	House C
Age	26	28	57
Gender	Male	Male	Male
Setting	Apartment	Apartment	House
Rooms	3	2	6
Duration	25 days	13 days	18 days
Sensors	14	23	21
Recorded by	Authors	Authors	Authors

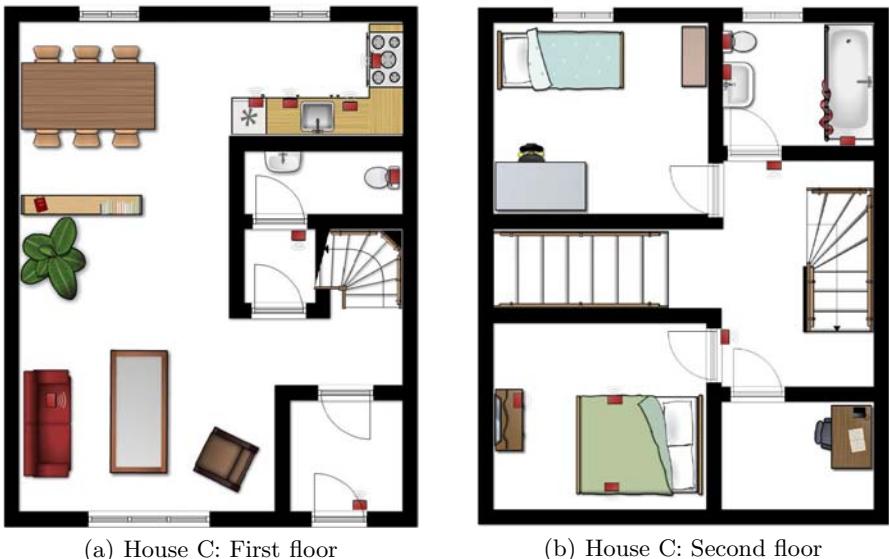


Fig. 3. Floorplan of house C, the red boxes represent wireless sensor nodes. (created using: <http://www.floorplanner.com/>)

and can be found in Table 4. Timeslices for which no annotation is available are collected in a separate activity labelled as ‘other activity’.

4.3 Experimental Setup

In all experiments the HMM was used as activity recognition model. All mappings that are performed use the meta feature list shown in Table 1, as discussed in section 3.1. Sensor data is discretized in timeslices of length $\Delta t = 60$ seconds. This time slice length is long enough to provide a discriminative sensor pattern and short enough to provide high resolution labelling results. After discretization we have a total of 35486 timeslices for house A, 17350 timeslices for house B and 26236 timeslices for house C.

We split our data into a test and training set using a ‘leave one day out’ approach. In this approach, one full day of sensor readings is used for testing and the remaining days are, depending on the experiment, either partly or fully used for training. We use each day as a test day once and report the average of the performance measure.

We evaluate the performance of our models using the F-measure, which is calculated from the precision and recall scores. We are dealing with a multi-class classification problem and therefore define the notions of true positive (TP), false negatives (FN) and false positives (FP) for each class separately as shown in a confusion matrix in Table 5, where N is the total number of classes.

Table 4. The activities that were annotated in the different houses. The ‘Num.’ column shows the number of times the activity occurs in the dataset. The ‘Time’ column shows the percentage of time the activity takes up in the dataset. All unannotated timeslices were collected in a single ‘Other’ activity.

Activity	House A		House B		House C	
	Num.	Time	Num.	Time	Num.	Time
Leave house	33	50.5%	16	50.6%	47	45.7%
Toileting	114	1.0%	28	0.6%	89	1.0%
Take shower	23	0.8%	8	0.6%	14	0.8%
Brush teeth	16	0.1%	12	0.2%	26	0.4%
Go to bed	24	33.2%	11	30.7%	19	29.2%
Prepare Breakfast	20	0.3%	7	0.5%	18	0.6%
Prepare Dinner	9	0.9%	2	0.2%	11	1.1%
Get drink	20	0.2%	11	0.2%	10	0.1%
Other	-	13.0%	-	16.4%	-	21.1%

$$\text{Precision} = \frac{1}{N} \sum_{i=1}^N \frac{TP_i}{TP_i + FP_i} \quad (2)$$

$$\text{Recall} = \frac{1}{N} \sum_{i=1}^N \frac{TP_i}{TP_i + FN_i} \quad (3)$$

$$\text{F-Measure} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (4)$$

4.4 Experiment 1: Meta-features vs. Sensor Feature Space

In this experiment we compare the performance of using the original sensor feature space to the performance of using the meta feature representation. We do not use any form of transfer learning in this experiment because it is not possible to do transfer learning using the original sensor feature space. Instead we train the model parameters using only data from a single house by performing maximum likelihood estimation, so no prior is used in learning the parameters.

Table 5. Confusion Matrix showing the true positives (TP), false negatives (FN) and false positives (FP) for each class. The ϵ_{ij} terms show the error between true class i and inferred class j . FN is the sum of the error terms in the same row, FP the sum of the error terms in the same column.

True	Inferred			FN
	1	2	3	
1	TP_1	ϵ_{12}	ϵ_{13}	FN_1
2	ϵ_{21}	TP_2	ϵ_{23}	FN_2
3	ϵ_{31}	ϵ_{31}	TP_3	FN_3
FP	FP_1	FP_2	FP_3	<i>Total</i>

In the case of the meta feature space the sensor data is mapped to the meta feature space. For the sensor feature space the sensor data can be used as it is. This allows us to do a fair comparison of the two feature spaces.

Figure 4 shows the results for all three houses. The X-axis shows the number of days of labelled data that was used, any remaining unlabelled data was also included during learning. The plot shows that the performance of the model using the meta feature space is slightly less than the model using the sensor feature space. This is because several sensors are combined into a single meta feature which gives the model less information for distinguishing activities.

4.5 Experiment 2: Separate Model vs. Single Model vs. No Transfer Learning

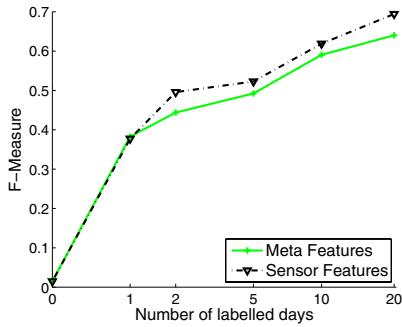
This experiment compares our transfer learning approach in which a separate model is used for each house with a transfer learning approach which uses a single model for all houses. A single house is used as target house, while the remaining two houses are used as source house. We compare the performance of these two transfer learning approaches to the performance of the model from the previous experiment in which no transfer learning and no mapping was used. This way we are able to see which transfer learning method works best and what the difference in performance is compared to not doing transfer learning.

Figure 5 shows the results for all three target houses. The X-axis show the number of days of labelled data that was used, any remaining unlabelled data was also included during learning. First of all we see that both our approach and the single model approach strongly outperform the ‘no transfer learning’ approach in all three houses when 0 days of labelled training data are used. This is because the ‘no transfer learning’ approach has no labelled data to learn its parameters, while the two transfer learning approaches can use the labelled data of the source houses.

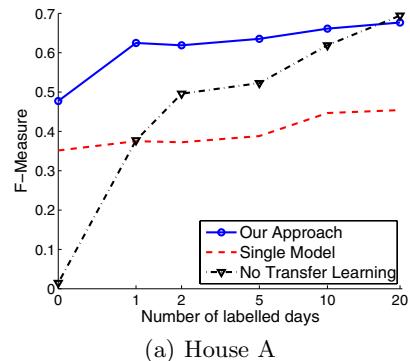
Furthermore, we see that our approach strongly outperforms the ‘single model’ approach in case of house A. This shows the benefit of having a separate model for each house, as can be seen from the jump in performance of going from 0 days of labelled data to 1 day of labelled data. Our approach is able to learn model parameters that take into account the specific behaviour for that house. The ‘single model’ on the other hand only gains a slight performance increase from this extra data, because it still shares the labelled data with the labelled data from the source houses. This makes the weight of the labelled data of the target house much less than when a prior is used.

Our approach also outperforms the ‘no transfer learning’ approach, although the difference in performance decreases as the number of labelled days increases. This clearly shows how the use of a prior helps in learning the model parameters and that its effect decreases as more labelled data is used.

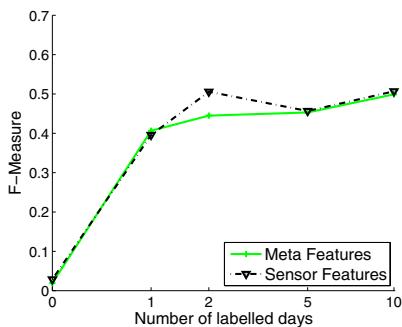
In the house B plot we see the ‘no transfer learning’ approach sometimes outperforms the transfer learning approach. This phenomenon is called negative transfer [5] which means that sometimes transfer learning can have a negative effect on the learning process. The reason for this is that it is not clearly defined



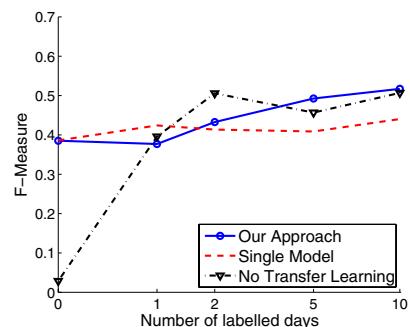
(a) House A



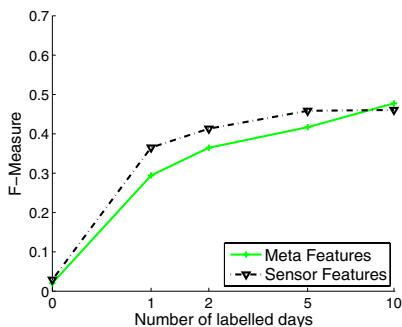
(a) House A



(b) House B



(b) House B



(c) House C

Fig. 4. Results of experiment 1, comparing the performance of the HMM using the meta-feature space and the original feature space. The x-axis are in log-scale and show the number of labelled days of data that were used for training.

Fig. 5. Results of experiment 2, comparing the performance of our transfer learning approach using a separate model for each house with a transfer learning approach using a single model for all houses and with an approach where no transfer learning is used. The x-axis are in log-scale and show the number of labelled days of data that were used for training.

which parts of data in the source houses are useful for the target house and which or not. Including training data from the source houses during learning can therefore sometimes pull the choice of parameters away from the optimal solution.

The house B plot also shows that our approach does slightly worse than the ‘single model’ approach when using 1 day of labelled data, but does better when more days of labelled data are included. This shows the advantage of using a prior, as more target data becomes available the learning method has to rely less on the prior (which caused the negative transfer). On the other hand, in the single model approach the data from both the source and target houses are all considered as valid training data for the single model. Therefore, a lot more target house data needs to be observed before it can outweigh the source house data, which is causing the negative transfer.

Finally, in the house C plot we see the ‘single model’ outperforms our approach when few days of labelled data are used, but as more days are added our approach manages to perform better or equal. This is similar to what we observed in the house B plot. Our approach slightly outperforms the ‘no transfer learning’ approach when a large number of labelled days is used.

4.6 Experiment 3: Labelled vs. Unlabelled Data

The use of generative models allows us to include unlabelled data during the learning process. In this experiment we compare performance of using both unlabelled and labelled data to using only labelled data. In both cases we use our transfer learning approach to learn the model parameters. The results for the various houses can be found in Figure 6.

We see that adding unlabelled data increases performance for house A, gives more or less equal performance for house B and decreases performance for house C, compared to the labelled only approach. The explanation of these mixed results is that the success of adding unlabelled data depends on the quality of the labelled data. We suspect that the use of a hand written diary for annotation (used in house C) results in less accurate annotation than using the bluetooth headset method (used in houses A and B). Although this less accurate annotation does not affect the learning process when using unlabelled data. It does affect the validation process when verifying if the inferred labels are correct.

4.7 Discussion

The results of our experiments show that our transfer learning method works well. Especially when no labelled data is available for a target house, our transfer learning approach is able to provide a good estimate of the parameters. But also in the presence of labelled data it can help in learning the model parameters. In some cases negative transfer can result in a lower performance compared to a non-transfer learning approach. This phenomenon has been reported in other transfer learning scenarios as well, but it is not well understood how to solve it [5].

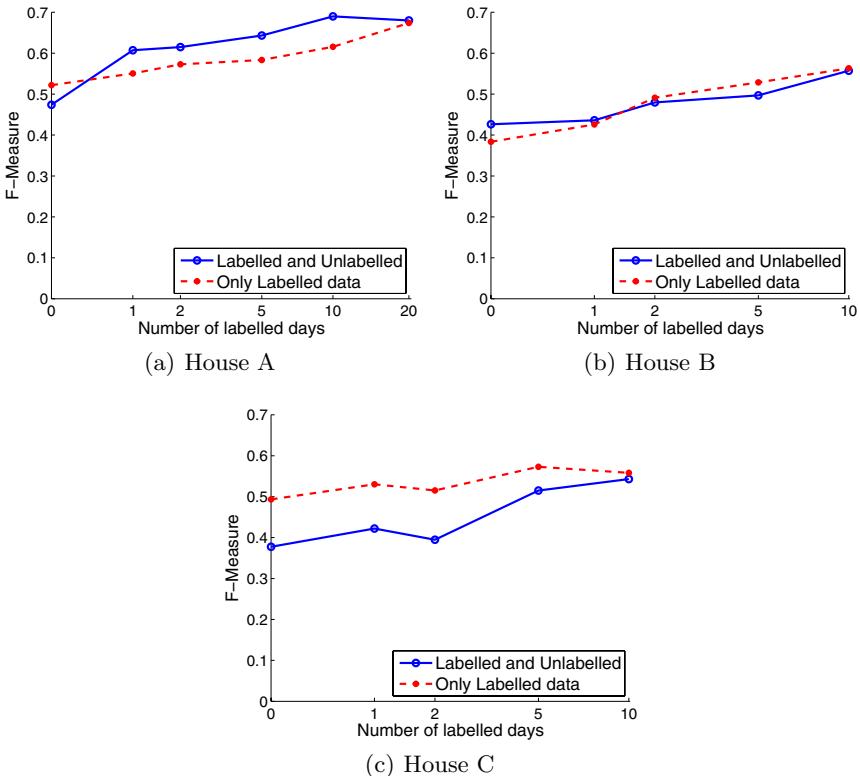


Fig. 6. Results of experiment 3, comparing the performance of using only labelled data with using both labelled and unlabelled data. The x-axis are in log-scale and show the number of labelled days of data that were used for training. In the case where unlabelled data is also included, the remaining days of unlabelled data are added during training.

An important factor in successfully applying transfer learning is the use of a proper mapping. In this work we manually defined the mapping beforehand. An alternative is to learn the mapping automatically from data [8,32]. However, because we are working with time series data trying various kinds of mappings might result in too high computation times for the approach to be feasible.

In terms of future work, it would be interesting to apply transfer learning to several other models. For example, the use of hierarchical models might be better fit for transfer learning because the different levels of the hierarchy allow a better abstraction between houses. Comparing the performance gain due to transfer learning between several models can provide interesting insights on how to accurately model data. It would also be interesting to apply our transfer learning approach to other sensing modalities such as camera's or wearables. Creating a proper mapping for those modalities will be challenging. Finally, it would be interesting to perform transfer learning across different sensing modalities. For example, using source houses in which camera's and wearables are used

to perform activity recognition and a target house in which a wireless sensor network is used.

5 Conclusion

We have addressed the problem of learning model parameters when little or no labelled data is available for the house activity recognition is to be performed in. Our main contribution is the introduction of a transfer learning method for activity recognition, which uses a prior to transfer general knowledge about activity recognition and allows the use of labelled and unlabelled data to learn house-specific behaviour.

Using experiments on three large real world datasets we showed our method gives good performance in activity recognition for a house for which little or no labelled data is available. The method can outperform a model trained using conventional maximum likelihood estimation. Furthermore, it can outperform a previously introduced transfer learning method in which a single model is used for all houses.

Acknowledgment

This work is part of the Context Awareness in Residence for Elders (CARE) project and the ‘Zorg(en) voor Morgen’ project. The CARE project is funded by the Centre for Intelligent Observation Systems (CIOS) which is a collaboration between UvA and TNO. The ‘Zorg(en) voor Morgen’ project is funded through the Pieken in de Delta-program by the Ministry of Economic Affairs and the cities of Utrecht and Lelystad and the provinces of Utrecht, Noord-Holland and Flevoland.

References

1. Abowd, G., Bobick, A., Essa, I., Mynatt, E., Rogers, W.: The aware home: Developing technologies for successful aging. In: Proceedings of AAAI Workshop and Automation as a Care Giver (2002)
2. Augusto, J.C., Nugent, C.D. (eds.): Designing Smart Homes, The Role of Artificial Intelligence. LNCS, vol. 4008. Springer, Heidelberg (2006)
3. Baxter, J.: A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning* 28(1), 7–39 (1997)
4. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, Heidelberg (2006)
5. Caruana, R.: Multitask learning. In: *Machine Learning*, pp. 41–75 (1997)
6. Chieu, H.L., Lee, W.S., Kaelbling, L.P.: Activity recognition from physiological data using conditional random fields. In: SMA Symposium. MIT Alliance, Singapore (2006)
7. Cook, D.J., Das, S.K.: Smart Environments: Technology, Protocols and Applications. Wiley-Interscience, Hoboken (2004)

8. Dai, W., Chen, Y., Xue, G.-R., Yang, Q., Yu, Y.: Translated learning: Transfer learning across different feature spaces. In: NIPS 2008: Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS 2008), Vancouver, Canada (2008)
9. Dai, W., Xue, G.-R., Yang, Q., Yu, Y.: Transferring naive bayes classifiers for text classification. In: AAAI, pp. 540–545 (2007)
10. Duong, T., Phung, D., Bui, H., Venkatesh, S.: Efficient duration and hierarchical modeling for human activity recognition. *Artif. Intell.* 173(7-8), 830–856 (2009)
11. Huynh, T., Schiele, B.: Towards less supervision in activity recognition from wearable sensors. In: Proceedings of the 10th IEEE International Symposium on Wearable Computing (ISWC), Montreux, Switzerland (October 2006)
12. Huynh, T., Schiele, B.: Towards less supervision in activity recognition from wearable sensors. In: ISWC, pp. 3–10 (2006)
13. Katz, S.: Assessing self-maintenance: Activities of daily living, mobility, and instrumental activities of daily living. *J. Am. Geriatrics Soc.* 31(12), 721–726 (1983)
14. Lee, S.-I., Chatalbashev, V., Vickrey, D., Koller, D.: Learning a meta-level prior for feature relevance from multiple related tasks. In: ICML 2007: Proceedings of the 24th international conference on Machine learning, pp. 489–496. ACM, New York (2007)
15. Lester, J., Choudhury, T., Kern, N., Borriello, G., Hannaford, B.: A hybrid discriminative/generative approach for modeling human activities. In: IJCAI, pp. 766–772 (2005)
16. Liao, L., Fox, D., Kautz, H.: Extracting places and activities from gps traces using hierarchical conditional random fields. *The International Journal of Robotics Research* 26(1), 119–134 (2007)
17. Minka, T.P.: Estimating a dirichlet distribution. Technical report, Microsoft Research (2000)
18. Oliver, N., Garg, A., Horvitz, E.: Layered representations for learning and inferring office activity from multiple sensory channels. *Comput. Vis. Image Underst.* 96(2), 163–180 (2004)
19. Patterson, D.J., Fox, D., Kautz, H.A., Philipose, M.: Fine-grained activity recognition by aggregating abstract object usage. In: ISWC, pp. 44–51. IEEE Computer Society, Los Alamitos (2005)
20. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2), 257–286 (1989)
21. Raina, R., Ng, A.Y., Koller, D.: Constructing informative priors using transfer learning. In: ICML 2006: Proceedings of the 23rd international conference on Machine learning, pp. 713–720. ACM Press, New York (2006)
22. Suzuki, R., Ogawa, M., Otake, S., Izutsu, T., Tobimatsu, Y., Izumi, S.-I., Iwaya, T.: Analysis of activities of daily living in elderly people living alone. *Telemedicine* 10, 260 (2004)
23. Tapia, E.M., Intille, S.S., Larson, K.: Activity recognition in the home using simple and ubiquitous sensors. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 158–175. Springer, Heidelberg (2004)
24. Thrun, S.: Is learning the nth thing any easier than learning the first? *Advances in Neural Information Processing Systems* 8, 640–646 (1996)

25. Truyen, T.T., Phung, D.Q., Bui, H.H., Venkatesh, S.: Hierarchical semi-markov conditional random fields for recursive sequential data. In: Neural Information Processing Systems, NIPS (2008)
26. van Kasteren, T., Englebienne, G., Kröse, B.: Recognizing activities in multiple contexts using transfer learning. In: Proceedings of the AAAI Fall Symposium on AI in Eldercare: New Solutions to Old Problems. AAAI Press, Menlo Park (2008), ISBN=978-1-57735-394-2
27. van Kasteren, T., Noulas, A., Englebienne, G., Kröse, B.: Accurate activity recognition in a home setting. In: UbiComp 2008: Proceedings of the 10th international conference on Ubiquitous computing, pp. 1–9. ACM, New York (2008)
28. Williams, M.T.: Beta-binomial distribution for proportional confidence intervals. Technical report, University of Leeds (1998)
29. Wilson, D.H.: Assistive Intelligent Environments for Automatic Health Monitoring. PhD thesis, Carnegie Mellon University (2005)
30. Wu, J., Osuntogun, A., Choudhury, T., Philipose, M., Rehg, J.M.: A scalable approach to activity recognition based on object use. In: ICCV, pp. 1–8 (2007)
31. Wyatt, D., Philipose, M., Choudhury, T.: Unsupervised activity recognition using automatically mined common sense. AAAI, 21–27 (2005)
32. Zhang, J., Ghahramani, Z., Yang, Y.: Learning multiple related tasks using latent independent component analysis. In: Weiss, Y., Schölkopf, B., Platt, J. (eds.) Advances in Neural Information Processing Systems, vol. 18, pp. 1585–1592. MIT Press, Cambridge (2006)

A Probability Distributions Used in the Hidden Markov Model

The HMM factorizes the joint probability over the observations and activities as

$$p(y_{1:T}, \mathbf{x}_{1:T}) = p(y_1) \prod_{t=2}^T p(y_t | y_{t-1}) \prod_{t=1}^T p(\mathbf{x}_t | y_t). \quad (5)$$

The individual factors are distributed as

$$p(y_1) \equiv \prod_{i=1}^K \pi_i^{\delta(y_1 - i)} \quad (6)$$

$$p(y_t | y_{t-1} = i) \equiv \prod_{j=1}^K a_{ij}^{\delta(y_t - j)} \quad (7)$$

$$p(\mathbf{x}_t | y_t) = \prod_{n=1}^N p(x_t^n | y_t) \quad (8)$$

$$(x_t^n = v | y_t = i) = (\mu_{in})^v (1 - \mu_{in})^{1-v} \quad (9)$$

where $\delta(x)$ is the dirac delta function giving 1 if $x = 0$ and 0 otherwise.

We use conjugate priors for the model distributions of the HMM.

$$\text{Dir}(\pi \mid \eta) = \frac{\Gamma(\sum_{k=1}^K \eta_k)}{\Gamma(\eta_1) \dots \Gamma(\eta_K)} \prod_{k=1}^K \pi_k^{\eta_k - 1} \quad (10)$$

$$\text{Dir}(a_i \mid \rho) = \frac{\Gamma(\sum_{k=1}^K \rho_k)}{\Gamma(\rho_1) \dots \Gamma(\rho_K)} \prod_{k=1}^K a_{ik}^{\rho_k - 1} \quad (11)$$

$$\text{Beta}(\mu_{in} \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu_{in}^{\alpha-1} (1 - \mu_{in})^{\beta-1}. \quad (12)$$

where $\Gamma(\cdot)$ is the gamma function. The parameters α and β are further parameterized as $\alpha_{in} = v_i^T f_n$ and $\beta_{in} = \omega_i^T f_n$, where f_n is a row vector of binary meta-features as shown in Figure 1. The hyperparameters v and ω are positioned in the meta feature space.

An overview of the probability distributions and their parameters is given in Table 2.

B Estimating the Hyperparameters

The maximum likelihood estimates of the parameters of the prior distributions cannot be found in closed form. We use numerical methods for estimating these parameters [17,28]. This gives us the values of α and β which are needed to find the values of the meta feature parameters v and ω . Because f_n is given we can find the least square solution to v and ω by solving the system of equations as defined by $\alpha_{in} = v_i^T f_n$ and $\beta_{in} = \omega_i^T f_n$. To guarantee a non-negative value we add a ‘bias’ meta-feature with a large enough positive value.

C Learning the Model Parameters Using the Prior

The MAP estimates of the HMM parameters can be found in closed form solutions by taking the derivative of the expectation function with respect to the parameter of interest. By using lagrange multipliers we can constrain the maximization to satisfy the rules of probability.

$$\pi_i = \frac{p(y_1 = i \mid x_{1:T}, \theta_{old}) + (\eta_i - 1)}{\sum_{i \in y_1} \{p(y_1 = i \mid x_{1:T}, \theta_{old}) + (\eta_i - 1)\}} \quad (13)$$

$$a_{ij} = \frac{\sum_{t=2}^T p(y_t = j, y_{t-1} = i \mid x_{1:T}, \theta_{old}) + (\rho_{ij} - 1)}{\sum_{t=2}^T \sum_{j \in y_t} p(y_t = j, y_{t-1} = i \mid x_{1:T}, \theta_{old}) + (\rho_{ij} - 1)} \quad (14)$$

$$\mu_{in} = \frac{(\alpha_{in} - 1) + \sum_{t=1}^T \xi_{inv} v_t}{(\alpha_{in} + \beta_{in} - 2) + \sum_{t=1}^T \xi_{inv}} \quad (15)$$

where $\alpha_{in} = v_i^T f_n$ and $\beta_{in} = \omega_i^T f_n$.

Common Sense Community: Scaffolding Mobile Sensing and Analysis for Novice Users

Wesley Willett¹, Paul Aoki², Neil Kumar¹,
Sushmita Subramanian², and Allison Woodruff²

¹ Computer Science Division, University of California, Berkeley, Berkeley, CA 94708 USA

² Intel Labs Berkeley, 2150 Shattuck Ave, Ste. 1300, Berkeley, CA 94704 USA

willettw@eecs.berkeley.edu, aoki@acm.org,

neilkumar@berkeley.edu, sushmita.subramanian@intel.com,

woodruff@acm.org

Abstract. As sensing technologies become increasingly distributed and democratized, citizens and novice users are becoming responsible for the kinds of data collection and analysis that have traditionally been the purview of professional scientists and analysts. Leveraging this citizen engagement effectively, however, requires not only tools for sensing and data collection but also mechanisms for understanding and utilizing input from both novice and expert stakeholders. When successful, this process can result in actionable findings that leverage and engage community members and build on their experiences and observations. We explored this process of knowledge production through several dozen interviews with novice community members, scientists, and regulators as part of the design of a mobile air quality monitoring system. From these interviews, we derived design principles and a framework for describing data collection and knowledge generation in citizen science settings, culminating in the user-centered design of a system for community analysis of air quality data. Unlike prior systems, ours breaks analysis tasks into discrete mini-applications designed to facilitate and scaffold novice contributions. An evaluation we conducted with community members in an area with air quality concerns indicates that these mini-applications help participants identify relevant phenomena and generate local knowledge contributions.

Keywords: Air quality monitoring, citizen science, environmental science, mobile sensing, participatory sensing, qualitative studies.

1 Introduction

Due to the increased availability of sensing technologies, citizens and novice users have new opportunities to pursue the kinds of data collection and analysis that were once handled almost exclusively by professional scientists and analysts [5]. Leveraging this citizen engagement effectively, however, requires not only tools for data collection but also mechanisms for understanding and utilizing citizens’ “local knowledge” – the experiential and cultural context, insights, and expertise unearthed through collaboration between locals and experts [4]. For example, while sensing



Fig. 1. A personal air quality sensor (left). Community members with sensors (right).

systems may be able to detect the presence of a pollution source, local insight may be required to actually identify the source or reveal sensitive populations affected by it.

Currently, most tools for viewing and analyzing sensed data do not explicitly support collaboration and are not designed to elicit or compile these kinds of local questions and insights. Moreover, analysis tools are generally not accessible to novice users, since they tend to assume a high level of technical and scientific literacy. We seek to understand how interactive systems for supporting citizen science can facilitate input from novice users and provide scaffolding that allows them to make greater local knowledge contributions.

This research is one component of the Common Sense project [1][8], a mobile sensing program that aims to deploy distributed air quality sensors in the service of practical action. Whereas traditional air quality monitoring organizations utilize coarse, representative measurements from a relatively small network of fixed sensors, we advocate a complementary mobile participatory sensing [3] approach in which large numbers of personal, mobile sensors are deployed within communities. This approach allows the community members impacted by poor air quality to engage in the process of locating pollution sources and exploring local variations in air quality. It leverages citizens' desire to understand personal exposure and knowledge of their communities to help effect change. We have developed a research testbed to explore this approach, examining issues such as the relative accuracy and resolution of community-sensed data versus data collected in professional fixed installations. The project also focuses on developing models for facilitating engagement and cooperation between community members, citizen scientists, activists, and other stakeholders.

In this paper, we survey related work in citizen sensing, collaborative visual analysis, and air quality presentation. We then discuss our own research, focusing on four key contributions: First, we present principles for designing for novice users in a citizen science setting, based on the results of extensive interviews with community members and other stakeholders in the air quality ecosystem. Second, we propose a framework for describing the process of local knowledge creation in citizen science. Third, we demonstrate the Common Sense Community site, a set of collaborative web-based visual analysis tools designed to facilitate collaborative analysis of sensed data and the co-production of local knowledge. Unlike prior systems, ours breaks analysis tasks into discrete mini-applications designed to facilitate and scaffold novice contributions. Finally, we present an evaluation of an early prototype of the site that

indicates these mini-applications help participants identify relevant phenomena and harness local insights.

2 Related Work

Citizen science and community environmental monitoring efforts have a deep and varied history that has been well documented in the environmental justice literature, illustrated by numerous examples of “backpack studies” and volunteer monitoring programs [4]. These examples have demonstrated the effectiveness of community participation in the collection of environmental data. O’Rourke and Macey discuss the use of “bucket brigade” sampling in which a mix of participants in different roles coordinate to carry out observation, sampling, and analysis of refinery emissions [25]. Other work has documented the use of community air quality sensing to identify polluters and enforce standards for diesel bus emissions [21][19]. This citizen-centric ethos has also begun to surface in government monitoring programs for water quality and waste [11].

Interactive tools for collaborative visual analysis may help community members and experts analyze community-sensed data, but the design of these tools presents numerous challenges [13]. Web-based tools like sense.us [14] and Many Eyes [30] have sought to facilitate collaboration using free-text comments attached to visualizations. However, this work has typically addressed short-term exploratory analysis of small datasets, rather than the long-term, iterative analysis associated with environmental monitoring. Meanwhile, commercial products for collaborative visual analytics [27][28] are targeted at expert analysts and are not accessible to novice users. Luther et al.’s Pathfinder [20] is perhaps the closest to our work. It seeks to utilize collaboration and visualization tools to support citizen science, but focuses on small datasets and wiki-based collaboration.

A number of projects have mapped air quality data using mobile sensors, typically with an emphasis on improving environmental awareness [9][17][26]. Some have taken creative approaches to presenting and collecting this data through artful visual presentation [6], provocative platforms [7], and gameplay [2]. While we also provide web-based tools to visualize data from mobile sensors, our tools focus instead on facilitating more direct engagement in the process of data analysis.

3 Motivating Fieldwork

Before deploying our mobile sensing platform with community members, we wanted to understand how those members factor into discussions about air quality and what roles they could play in data collection, analysis, and outreach. To gauge this, we conducted a concentrated investigation of the communities we hoped to engage with.

3.1 Method

Over the course of several months, we interviewed novice community members as well as scientists, remediation consultants, government representatives and other stakeholders in order to understand their perspectives on air quality and assess the role

that technological interventions could play in their environmental decision-making processes [1]. This included 14 formal, in-person interviews and approximately 30 informal interviews conducted either in person, by phone, or at community meetings. In these interviews, we discussed existing practice and used prototype sensors and interface mockups to explore people's reactions to potential mobile sensing tools. We recorded the formal interviews and took detailed field notes describing all of our interactions. Using these, we performed affinity clustering to identify a general set of emergent themes and design principles. We also performed more targeted clustering to identify common user needs, tasks, and motivations for community participation and engagement with environmental data.

3.2 Personas

Based on this fieldwork, we developed a set of personas to characterize the relevant stakeholders and identified a set of common tasks and questions associated with each. Because the system presented here is targeted primarily at community members and novice users, we will limit our discussion to the three most relevant personas: an *activist* or *community organizer* responsible for orchestrating actions and publicizing environmental issues, a *browser* who has an interest in environmental quality but is not directly involved with sensing, and a novice community member who might act as a *data collector* (Table 1). While we focus here on tools for these community members and novice users, it is also clearly valuable to provide tools for (and promote dialog with) other expert stakeholders with different needs, such as *scientists* and *government regulators*.

Table 1. Some of the key personas derived from our initial fieldwork

	Activist/Organizer	Browser	Data Collector
Motivation	Specific concerns about the community with an emphasis on political change.	Likely to be interested in environmental and/or societal issues. Possibly concerned with political change.	Likely to have personal health issues.
Goals	Prove there is a problem. Determine neighborhood exposure. Pursue political change.	Understand broader environmental and societal impacts. See trends.	See personal, immediate data. Modify personal behavior. Pursue political change.
Desired Tools	Tool for community understanding and presentation.	Summaries, Interactive tools for exploring data.	Glanceable summaries, Alarms, Forecasting.

3.3 Design Principles

Based on our fieldwork, we also extracted a set of design principles for developing tools to support visual analysis of sensed data. Some of the key issues are:

Support specific, goal-directed tasks. Participants were highly goal-oriented and motivated by specific issues such as "What is my personal exposure throughout the

day?” or “What are hotspots in this area?”. “General” exploration did not tend to engage them. As one interviewee put it, “You don’t want to look at the interface and say, ‘What is this supposed to tell me?’”

Show local and personally relevant data. Participants were most interested in data close to their homes and other locations they frequented, rather than the aggregate regional data typically provided by current air quality monitoring solutions. The interviews further suggest that many users may not engage unless they are driven by health concerns or some other issue that personally connects them to the data. As one participant said, “Make the data as local as possible. People want to see their house, their block, not a general neighborhood, not a general area.”

Elicit latent explanations and expectations. Community members have local knowledge and expertise, such as beliefs about sources of pollution in their neighborhood. However, our interviews suggest that it is often difficult for them to translate this knowledge into specific queries. While community members were good at generating high-level or vague questions (e.g. “How does the freeway impact air quality?”), they had fewer immediate instincts about how to break these questions down. Therefore, it is important to provide tools that help community members draw on their personal knowledge, for example by making suggestions about possible formulations of queries or by guiding them in their exploration of the data.

Prompt realizations. As mentioned above, community members have significant local knowledge that could be helpful in interpreting local environmental data. Accordingly, it is valuable to present views of the data that are perceptually suggestive of various possible patterns, and therefore prompt spontaneous realizations that draw on the users’ local knowledge. For example, a view that aligns readings from multiple days may prompt a user to realize that repeated spikes at a site are the result of a recurring event – for example, a delivery truck unloading.

Beware of “language” barriers. Current tools to which community members have access, such as the EPA EnviroMapper [11], are technically complex and require a moderate level of scientific knowledge (for example an understanding of pollutant concentrations in parts per million). Novice users may benefit from scaffolding to introduce scientific language, and tools that target novice users should not require an understanding of such language.

“You don’t want to be inundated.” Understandably, participants did not want to be overwhelmed with unnecessary information and complexity (particularly if the information was somewhat new to them or was beyond their level of expertise). Therefore, staged or gradual presentation of information is desirable.

3.4 Framework

Drawing on our personas and design principles, we derived a framework for describing data collection and local knowledge generation in a citizen science setting. This framework does not just describe the existing ecosystem or citizen science applications. Rather, it builds on the key findings and user needs we identified in our

fieldwork and describes operations an ideal citizen science solution might address. As such, the framework serves as a potential blueprint for designing new citizen science tools and for assessing existing ones.

In this framework, we divide the process of collecting, analyzing, and synthesizing environmental data and local insights into six phases: *collect*, *annotate*, *question/observe*, *predict/infer*, *validate*, and *synthesize*. While these phases can build on one another, they are not necessarily linear and individual participants do not necessarily participate in all of them. Rather, each involved stakeholder may engage in the process at a few phases and the various members of the community together carry out activities at all phases. The various phases each serve different functions and can build on one another but do not always do so. These phases may also be iterative - for example, answering questions and validating predictions may require additional data collection.

The phases detailed here dovetail with formulations of the scientific method, and some steps (*question*, *predict*, and *validate*) echo the question-hypothesize-test formulations seen in the science education literature. However, our framework describes a more general set of operations, many of which need not necessarily be formulated in the language of scientific discourse. Questions, predictions, and inferences generated by community members are often pre-scientific and can contribute valuable insights that inform a more formal and rigorous process of scientific analysis without necessarily being framed as such.

Finally, while we frame this process in terms of air quality monitoring for the sake of this discussion, the framework itself is applicable to a broad range of citizen science projects including other environmental and health monitoring efforts.

Collect

In this phase, data collectors engage in various collection activities. These may include using sensors to record raw data or observing phenomena and making manual observations (as in traditional citizen science activities like the Christmas Bird Count [24]). Most existing citizen science places a strong emphasis on this collect phase.

Annotate

After data has been recorded, data collectors provide additional insights that contextualize and supplement it. This can include additional information that helps explain the data; for example, if a peak in the data corresponds to an event they observed during collection. Collectors can also include information about the data gathering process (when, where, and under what conditions was the data collected) or comments about data quality.

Question/Observe

Using their own data and data collected by other participants, data collectors (as well as browsers and activists) can begin to ask basic questions and identify trends. These questions can be introspective (“What is my personal exposure to pollutants?”, “Is air quality bad at my home?”) or generally inquisitive (“Where is air quality good and bad?”, “Are there block-by-block trends in air quality?”). Some of these questions, including those dealing with personal exposure, can often be answered directly using the collected data, while others are more abstract. These questions can be implicit or

explicit and may be driven by the data or by existing assumptions and expectations. Users may also observe and note apparent trends (for example, higher levels of a pollutant at different times of day) or other phenomena of interest (high levels at an unexpected intersection).

Infer/Predict

Building on these questions and observations, data collectors, browsers, and activists can begin to make predictions and inferences about the observed phenomena (“I think values will get worse towards this intersection.”, “Higher readings here seem to indicate a source.”). The observations and inferences made by community members may be less clearly articulated than in a formal analysis, but can contain local insights. While this phase often resembles the “hypothesize” stage seen in formulations of the scientific method, participants’ predictions and insights may not necessarily be framed as clearly testable hypotheses. They may only suggest the existence of a trend or its repeatability rather than proposing a mechanism for it. In these predictions, regardless of their precise formulation, lie some of the most important pieces of local knowledge that community members can contribute.

Validate

At this phase, contributions from data collectors are more likely to overlap with those of activists and organizers. Here, data collectors, browsers, and organizers may look for additional data to corroborate their own findings and organizers may also make requests for additional data. Additionally, organizers may enlist the help of outside entities including domain experts and professional analysts to help verify insights and predictions generated by collectors and browsers.

Synthesize

At the highest level, activists and organizers must integrate the data and knowledge generated in prior phases to produce documentation, reports and other deliverables. Again, organizers may involve domain experts and professional analysts, along with administrators and regulators, in order to generate summary documentation that can be used to support activism, inform policy decisions, and enforce regulations.

This framework (and particularly the *annotate*, *question/observe*, and *infer/predict* phases) provides a blueprint for scaffolding novice users’ progression from initial elicitation through more involved and integrated questions and contributions. In this paper, we focus on applications that engage novice users and guide them through these initial phases. We defer discussion of validation and synthesis, which tend to utilize more specialized sets of tools for more expert users.

4 The Common Sense Community Site

Building on the framework and our design principles, we designed and built the Common Sense Community site, a suite of task-oriented mini-applications that allow community members to participate in the collaborative analysis of local air quality data. While the site is targeted primarily at novice data collectors in a low-income urban area, it is also designed to be accessible to more specialized participants

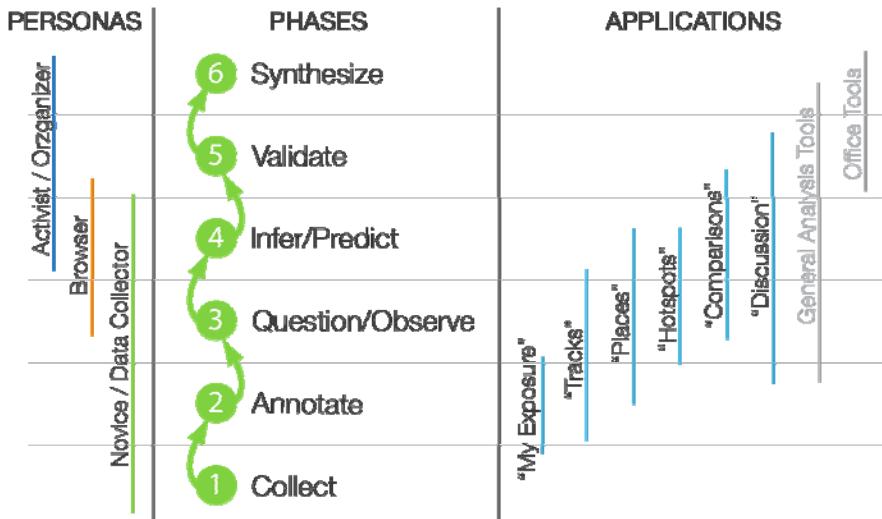


Fig. 2. Our framework for knowledge generation in citizen science (center). Personas (left) and tools (right) are shown in their intended phases.

(browsers, organizers, scientists, administrators, and regulators) who may engage in the analytic process at different phases.

The set of visualizations is designed specifically to facilitate the incremental progression of novice community members through multiple phases of analysis. A person may begin by collecting data or asking questions about data collected by other community members and progress through structured phases, triggering new kinds of insights. Over time, this can allow novices to become more adept contributors.

Providing a suite of simple task-oriented applications rather than a more general analysis tool has several benefits. First, it lowers barriers to entry. Participants do not need to learn a complicated tool in order to contribute. This encourages legitimate peripheral participation [18] and allows novice users and participants with little computing experience to engage in the process. Whereas more general analysis tools such as Excel, Tableau [27], or Matlab require greater familiarity with formal analysis processes, these individual applications allow users to answer specific questions and can guide them towards particular kinds of insights. Figure 2 shows approximate mapping between our mini-applications and the framework discussed previously.

4.1 Collecting Data

Users collect air quality data using mobile sensors designed as part of the broader Common Sense project [8]. These sensors (Figure 1) are designed to be self-contained and unobtrusive monitoring devices that can be clipped to a bag or carried as an accessory. The units feature a custom board design and embedded software that can be deployed with commercial carbon monoxide, nitrogen oxides, and ozone gas sensors. As users carry these sensors with them throughout the day, the units transmit

live sensor reading and GPS data to a database server over a GSM data network connection. Users can also upload data from offline air quality sensors.

4.2 Applications

To display this data, we built mini-visual analysis applications that target common, representative tasks and questions that we identified through our fieldwork. These included: monitoring personal exposure, inspecting recorded tracks, identifying locations with poor air quality, and eliciting possible sources. These targeted applications exemplify our approach to designing for citizen science – modular, accessible applications that serve specific needs and which together scaffold the process of local knowledge production. Users begin by selecting an application that serves a particular need (e.g. “see my personal exposure”) from a portal site. They then move between applications via a tabbed interface. We also provide gateways designed to allow participants to build familiarity with simpler, more targeted tools and then transition in a natural way to more complex tools designed to elicit different types of insights. This facilitates the transitions between *annotation* and *questioning* or *questioning* and *inference* we described in our framework.

In each of these applications, users can record their questions and insights by leaving comments attached to individual views of data. Each application features a commenting panel (Figure 4c) that participants can use to annotate and discuss their findings. This panel also provides intelligent prompts designed to elicit questions and observations, along with educational prompts designed to help scaffold novice users’ understanding of the domain.

We describe several applications in detail below.

My Exposure

The first application provides a widget that helps users answer one of the most common questions we observed in our fieldwork: “What is *my exposure* to a pollutant?” Many of the community members we interviewed suffered from allergies or respiratory disease exacerbated by the poor air quality in their neighborhood, and expressed a desire for tools that would help them gauge and mitigate their exposure. To meet this need, we developed the *My Exposure* widget (Figure 3, Figure 4a). *My Exposure* shows a single aggregated measure of the pollutants measured by a participant’s sensor, normalized over time to the EPA’s Air Quality Index (AQI) [22]. (Because many people are not familiar with raw pollutant concentrations, all of the visualizations on the site also use the AQI color encodings and category descriptors – “Good”, “Moderate”, “Unhealthy for Sensitive Groups”, “Unhealthy”, “Very Unhealthy”, and Hazardous” – in addition to providing actual values).

For community members carrying our air quality sensors, this application acts as an entry point to the site and serves an ongoing need that is likely to garner repeat



Fig. 3. Two views of the *My Exposure* application



Fig. 4. The Common Sense Community Site showing data collected by a single user. The *My Exposure* widget (a) and *Tracks* visualization (b) are visible along with the commenting panel (c).

visits. To encourage participants who are initially only curious about their exposure to further explore their data, we placed the *My Exposure* view adjacent to the *Tracks* application (discussed momentarily).

Tracks

The *Tracks* application (Figure 4b) provides a simple way for novice users to observe and ask questions about pollution data from their own sensor. In this visualization, pollution measurements are plotted on a map and also appear in a timeline below the map view. The application behaves like a media player and provides a play/pause button, a playback speed control, and a draggable thumb on the timeline that can be used to scrub back and forth in the dataset.

As mentioned above, in each of our applications, participants use the commenting panel (Figure 4c) to annotate and discuss their findings. This panel is collapsed by default to avoid overwhelming the user, but expands to display intelligent prompts designed to elicit questions and observations. For example, when a participant plays back data from their own sensor in the *Tracks* application, the interface pauses briefly whenever a dramatic spike occurs in the data and actively prompts the user to document the change. The user can choose to either enter a comment or continue playback. If no action is taken, playback resumes after a brief interval. Users can also pause playback at any point to enter comments or questions.

Places

Our fieldwork indicated that users' initial inquiries about air quality are often location-centric ("What is air quality like in *my* neighborhood?", "Are we protecting our 'treasures', our schools, hospitals, libraries, parks, etc.?"). To help facilitate questions and observations of this type, we provide a location-centric *Places* visualization

(not pictured). When a user starts the visualization, they are prompted to enter an address and a time range. The application then produces an interactive map showing all data collected by any sensor near the specified address during those times. Whereas the *Tracks* application is designed to mimic the functionality of a media player, *Places* is designed to feel similar to online mapping tools like Google Maps [12]. The map can be panned and zoomed and the data points plotted on it can be played back chronologically.

We include gateways that allow users to enter the *Places* view from within other applications. When using another application, a user can click a “see more for this location” button to transition to the *Places* view, centered on the location visible in their current application.

Hotspots

The *Hotspots* visualization (not shown) helps users identify regions with the best and worst air quality over a period of time. The application is intended to help users answer questions about where and when levels are high and low. It draws on the notion, frequently seen in our initial interviews, that “worse things are exciting” and uses this to provoke insights regarding new locations and unexpected sources.

Using a range slider, users select whether to show regions with high or low pollution levels. Readings that match the specified thresholds are then plotted on a map similar to the one used in the *Places* view. Users can also transition to this visualization by clicking the “see other places with readings this high/low” gateway from within the *Tracks* or *Places* applications.

Comparisons

The *Comparisons* visualization is designed to support inference and help users identify repeated sources and relationships between them. The *Comparisons* visualization presents users with a set of discrete ‘episodes’, short windows of time in which some notable event occurred in the recorded air quality data. These can be the largest spikes seen in an area over the course of a period of time, or the periods of time with the highest variance.

The notion of focusing on spikes was driven by two observations from our field-work. First, we noted that people often wanted to “examine an event, not a timeline,” seeing detailed data at the scale where the event was apparent, rather than at the level of the entire dataset. Second, we hoped that by grouping together sets of episodes that would otherwise appear separately, this view would prompt noticing and inferences that might not emerge otherwise. In the *Comparisons* view, these episodes are displayed as a set of small multiples [29] alongside a map that also plots that same data (Figure 5). The small multiples are linked to the map so that brushing a plot focuses that event in both views. This allows users to compare the events spatially as well as temporally.

Discussions

In addition to the collapsible commenting pane that accompanies each one of the visualizations, the site features a *Discussions* view – a separate application that serves as a central location for viewing all comments and provides a forum-like interface for

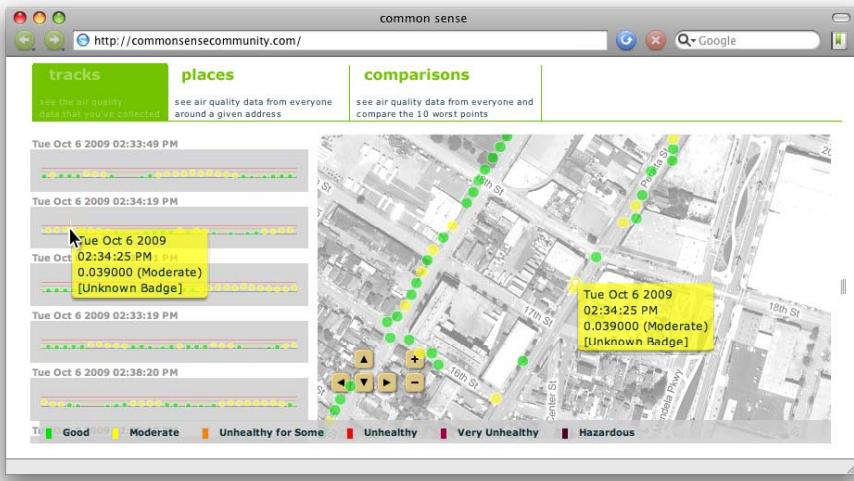


Fig. 5. The *Comparisons* view. Small multiples of the timeline (left) showing the five highest episodes recorded during the past day. The commenting panel is hidden at right.

further discussion. All comments and annotations left by users in the other applications are visible here as separate threads and users can compare and build on observations and insights from multiple applications.

Other Candidate Applications

The visualizations described here cover a large subset of the kinds of questions and observations specified by our framework. However, other visualizations are clearly possible (e.g., tools for understanding variations in air quality over time), and we expect to build examples of them in the future to support other relevant questions.

We also anticipate tools that will assist participants in the *validate* and *synthesize* phases of a citizen science task. For example, applications might include automated pattern matching to help locate sets of similar sources or identify characteristic pollution signatures. Similarly, tools to support “crowdsourcing” could allow organizers to request new samples or ask community members to identify sensitive locations like schools and day care centers.

4.3 Implementation Details

The Common Sense Community site and all of the visualizations within it were constructed using Adobe Flash with the Modest Maps toolkit [23].

5 Evaluation

We deployed an early version of the site with community members in a low-income urban neighborhood with poor air quality. There, we carried out interviews and think-aloud assessments to help characterize participants’ use of the tools. We wanted to

understand which visualizations were perceived to be useful and approachable and assess whether this set of tools facilitated activities at the various phases identified in our framework, such as emergent prediction and observation.

5.1 Method

During our assessment we carried out seven interviews with nine community members. We recruited participants through a local non-profit organization that focuses on environmental monitoring and awareness. Five of the participants were affiliated with the non-profit and had participated in air quality monitoring activities through the organization. Most of the participants we surveyed were members of a small and relatively tightly knit community and the majority knew one another in some capacity. Participant ranged in age from the mid-teens to late 40's and had a variety of education levels, including some middle- and high-school students and some participants without high school degrees.

We conducted all of the interviews at the office of the non-profit. We started each session with a brief interview designed to assess participants' knowledge of air quality issues and the impact of air quality on their community. In our discussions, we emphasized the impacts of particulate matter and described its sources. We then gave the participants a particulate matter sensor and asked them to take samples in a several block radius around the office. We asked participants to choose a route that they thought would maximize the amount of particulate matter detected. During the sampling process, the interviewer walked with the participants and asked them to describe their route choice and identify potential sources in the area. We used a commercial particulate matter sensor rather than our custom hardware since particle pollution is of particular interest in the target neighborhood.

Once they returned to the non-profit, participants used an early version of the Common Sense Community site to examine their data as well as data gathered by other participants. We conducted a one-hour think-aloud evaluation with each participant in which they were instructed to interact with the site and verbally relate their thought processes and any questions or insights that occurred to them. Participants used a version of the site that included the *Tracks*, *Places*, and *Comparisons* visualizations detailed above. In the *Places* and *Comparisons* views, each participant had access to his or her own measurements as well as measurements taken by all of the previous participants. Because users only had access to data collected by a small group of participants in short windows over the course of a few days, we were unable to test the *Hotspots* visualization, which was designed to leverage larger datasets.

We recorded each of these interviews and coded participants' interactions with the site to assess whether or not they fit within our framework. We also performed clustering to extract key findings that emerged. These are discussed in Section 6.

5.2 Scaffolding and Navigation Strategies

Most participants were able to explore the visualizations and inspect the data that they had collected without much confusion. The majority began by identifying their current location on the map and followed the track they had recorded, looking for peaks either on the map or in the timeline. Most voiced questions and observations about the

data and a few made additional inferences or predictions. We report key observations that correspond to each of the phases in our framework.

Collect. Almost all of the users identified a nearby freeway and trucking lots as the most likely sources of pollution and most chose routes that took them along a nearby frontage road. The students we interviewed all minded their sensors attentively as they walked, looking for spikes and actively seeking out areas with higher readings. All other participants used the sensor more passively and traversed areas that they predicted would be more polluted without actively noting the levels there.

Annotate. Using the Tracks view, several participants observed distinct peaks and ascribed them to events that occurred or features that they passed while they were walking (“All the trucks [get on the highway there].”, “That’s the new construction there.”). Participants also tended to note readings taken adjacent to locations that interested them (“At least we don’t have any red marks near the park...”). In two cases, participants had observed increased particulate matter levels on the sensor as they walked and directly attributed a peak to a particular source.

Question/Observe. Most participants asked questions and made remarks about locations (“Where was that again?”), data (“Was [that spike] at an intersection?”), and other participants (“Where did she go?”, “Which person did that come from?”). Participants also asked broader questions about day-to-day and month-to-month trends. For example, one wondered whether pollution levels would change during the rainy season and another asked “Would it be different if there was wind?” A few participants also noted locations on the map without data and contributed additional anecdotes and pieces of information about them.

Infer/Predict. Based on the data and their initial questions and observations, several participants made inferences about the behavior of phenomena they observed. For example, one participant compared her readings with those from a participant earlier in the day and noticed that her own were higher. She inferred that the level of particulate matter might be impacted by the change in temperature.

Another participant investigated the data he had collected and extrapolated from it to predict air quality readings further along the frontage road saying, “I wouldn’t doubt that it gets worse around the bend.” Talking about a several-block radius, he also made a prediction about the health impacts of pollutants in the area. He noted, “Just in this radius I can honestly say [...] at least half the kids have asthma. At *least* half.” He supplemented this prediction with a quick calculation, “Fifteen residences per area so ... that’s probably about a good 500 kids.”

Validate and Synthesize. This set of interviews involved only novice community members and incorporated only data collected during their sessions. As such, we did not emphasize the validate and synthesize phases in this study.

5.3 Usability

Based on our fieldwork, we were mindful in our design process of the computer literacy of the target population. As one participant in our initial interviews noted,

“There’s still that big digital divide in [our city] and all poor neighborhoods.” Therefore, we were pleased that the system was generally usable by all participants. The study did reveal a few straightforward usability issues, which we are addressing, such as the need to make the playback controls more visible. These issues did not appear to impact the results discussed below.

6 Discussion

Here we discuss trends and activities we observed across all of our interviews.

6.1 Health and Personal Safety

As expected, displays tailored to personal use proved to be an effective tool for engaging users in the process of citizen science. The most interested and receptive participants each had a personal or family health concern (asthma, allergies, or some other reaction) that they attributed to air quality. One asthmatic participant who bicycles and does not own a car expressed a desire to use the data to vet safe cycling routes, stating, “This has brought to mind – you’re gonna get exercise, but what are you breathing in?” Participants with small children also expressed a strong desire to use the tool on a regular basis to help minimize exposure.

6.2 Socializing

Although we conducted interviews separately and the sequential nature of the interviews did not facilitate conversations or dialogues using the commenting tools, we did see social interactions between participants when they viewed one another’s data. Several participants asked questions like, “Which person did these come from?” and “Whose was whose?” and were eager to compare their tracks against those recorded by previous participants. In particular, those from the same social circle were interested in knowing which of their friends had collected data, where they’d walked, and how “well” they had done. For example, one participant located a friend’s track and followed it for the entire length, noting each location she’d visited and commenting, “She was pretty good, [she found a few orange ones].” Comparing tracks in a competitive way was also common, particularly among the students we interviewed. One group of younger students, for example, was excited to discover that their readings were higher than those of other participants. This suggests a competitive impulse that we might also leverage to encourage participation.

During the interviews, several participants attributed their continued awareness and investment in air quality to a particular community organizer. One participant observed, “You could say *she’s* our resource when things are happening. If she feels we need to know, then it’s up to us to get involved.” This suggests that, at least within this community, maintaining long-term interest and investment depends, in part, on leveraging these kinds of key community members.

While we observed users’ reactions to one another’s data, the linear nature of our interviews did not allow us to observe exchanges or evolving social use of the system. A longitudinal study with more users is needed to understand these social aspects of

the system and to gauge the impact of larger amounts of data and discussion on the analyses that participants undertake.

6.3 Exposing Preconceived Notions

A number of our participants approached the data not from an inquisitive standpoint, but rather expecting to find validation of their expectations about air quality. We noted comments from a number of participants that suggested implicit assumptions about areas (“On Fourth Street, that makes sense.”) and expectations about how bad pollution levels would be (“[If you sampled this area] you’d see lots of red”). One participant, in particular, was surprised that the level of particulate matter she recorded was low, stating, “I feel like it should be a little stronger with picking up certain particulates and fumes. *I know* there should be a lot more out there because there are a lot of businesses and industrial stuff.” To test this, the participant requested to take the sensor out again and collected additional data.

In some cases these kinds of assumptions may function as implied hypotheses and predictions that participants can immediately begin to validate and build on. However, as in the case of the latter participant, preconceptions can sometimes generate mistrust in sensors and tools that do not reinforce these existing notions.

6.4 Visualizations as a Catalyst for Discussion

We also observed several participants who used the map extensively as a catalyst for discussion. These users would point and navigate to areas with strong personal relevance including their homes, schools, and public areas, even when no air quality data for that particular region was present.

One interviewee, in particular, used the map to discuss pollution sources outside the zone in which he had collected data and to make predictions about sources and impacts there. He first predicted that there might be “really high values” in main intersections adjacent to a nearby port and shipping terminal, stating, “I can only imagine [it gets worse toward the intersections.]” He then contributed a number of anecdotes about locations in and around the port including spots where diesel trucks idle, areas where water quality has been impacted by dredging, and an isolated residential building in the industrial zone. These anecdotes were often very specific and drew on his experience as a port worker and volunteer air monitor – for example:

“Here - definitely this intersection - we did some of the survey in this area last year. Here, right here - this is a fuel station. It’s a truck fuel station. This is where all the trucks get on the freeway. All the trucks are always right here - along [Street 1] and [Street 2] and um, [Street 3] and [Street 2]. I know for sure, these monitors are not going to catch moderate here. Lucky enough, nobody lives on these blocks. All business, all industry.”

These kinds of observations are key examples of the types of local insights community members may bring to the table and which we hope to elicit.

7 Conclusions and Future Work

In this paper, we have presented design principles for targeting novice users in a citizen science setting and supplied a framework that describes data collection and knowledge generation in these conditions. We have described the genesis of this model through interviews with community members and activists, as well as its application in the user-centered design of a system for mobile air quality monitoring. Unlike prior systems, ours breaks analysis tasks into discrete mini-applications designed to facilitate and scaffold novice contributions. Based on our initial evaluations, this strategy helps novice users identify relevant phenomena and generate local knowledge contributions.

Although the applications discussed here focus on air quality, we believe that the approach we advocate can be applied to other domains with a citizen science component. Monitoring of other environmental indicators including water and soil quality as well as epidemiological monitoring should be equally applicable, particularly when the object of study is of strong significance to the participants. As we move forward to deploy mobile sensors more broadly and develop mobile interfaces for accessing and interacting with the data, we expect to employ similar techniques and build on these frameworks and tools.

Acknowledgements

We gratefully acknowledge the support of Brian Beveridge, Margaret Gordon, and our other collaborators at the West Oakland Environmental Indicators Project. We also thank Rob Ennals and Lora Oehlberg for their helpful discussions and comments and acknowledge Ron Cohen, Prabal Dutta, RJ Honicky, Alan Mainwaring, Chris Myers, Eric Paulos, Paul Wooldridge, and our study participants for their valuable contributions to this work. Common Sense builds in part on the Participatory Urbanism [26] and N-SMARTS [15] projects.

References

1. Aoki, P.M., Honicky, R.J., Mainwaring, A., Myers, C., Paulos, E., Subramanian, S., Woodruff, A.: A Vehicle for Research: Using Street Sweepers to Explore the Landscape of Environmental Community Action. In: Proc. CHI 2009, pp. 375–384. ACM, New York (2009)
2. Black Cloud, <http://www.blackcloud.org>
3. Burke, J., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., Srivastava, M.B.: Participatory Sensing. In: SenSys 2006 WSW Wksp. ACM, New York (2006)
4. Corburn, J.: Street Science: Community Knowledge and Environmental Health Justice. MIT Press, Cambridge (2005)
5. Cuff, D., Hansen, M., Kang, J.: Urban Sensing: Out of the Woods. Comm. ACM 51(3), 24–33 (2008)
6. Da Costa, B., Schulte, J., Singer, B.: AIR – Area’s Immediate Reading, <http://www.pm-air.net>
7. Da Costa, B., Hazegh, C., Ponto, K.: PigeonBlog, <http://pigeonblog.mapyourcity.net>
8. Dutta, P., Aoki, P.M., Kumar, N., Mainwaring, A., Myers, C., Willett, W., Woodruff, A.: Common Sense: Participatory Urban Sensing Using a Network of Handheld Air Quality Monitors (demonstration). In: Proc. SenSys 2009, pp. 49–50. ACM, New York (2009)

9. Eisenman, S.B., Miluzzo, E., Lane, N.D., Peterson, R.A., Ahn, G.S., Campbell, A.T.: The BikeNet Mobile Sensing System for Cyclist Experience Mapping. In: Proc. SenSys 2007, pp. 87–101. ACM, New York (2007)
10. Environmental Protection Agency: Volunteer Monitoring | Monitoring and Assessing Water Quality, <http://www.epa.gov/volunteer/>
11. EPA EnviroMapper, <http://www.epa.gov/enviro/html/em/index.html>
12. Google Maps, <http://maps.google.com>
13. Heer, J., Agrawala, M.: Design Considerations for Collaborative Visual Analytics. *Information Visualization* 7(1), 49–62 (2008)
14. Heer, J., Viégas, F., Wattenberg, M.: Voyagers and Voyeurs: Supporting Asynchronous Collaborative Information Visualization. In: Proc. CHI 2007, pp. 1029–1038. ACM, New York (2007)
15. Honicky, R.J., Brewer, E., Paulos, E., White, R.: N-SMARTS: Networked Suite of Mobile Atmospheric Real-Time Sensors. In: SIGCOMM 2008 NSDR Wksp. ACM, New York (2008)
16. Irwin, A.: Citizen Science: A Study of People. In: Expertise and Sustainable Development. Routledge, London (1995)
17. Kanjo, E., Benford, S., Paxton, M., Chamberlain, A., Stanton Fraser, D., Woodgate, D., Crellin, D., Woolard, A.: MobGeoSen: Facilitating Personal Geosensor Data Collection and Visualization Using Mobile Phones. *Pers. Ubiquitous Computing* 12(8), 599–607 (2008)
18. Lave, J., Wenger, E.: Situated Learning: Legitimate Peripheral Participation. UP, Cambridge (1991)
19. Levy, J.I., Houseman, E.A., Spengler, J.D., Loh, P., Ryan, L.: Fine Particulate Matter and Polycyclic Aromatic Hydrocarbon Concentration Patterns in Roxbury, Massachusetts: A Community-Based GIS Analysis. *Environmental Health Perspectives* 109(4), 341–347 (2001)
20. Luther, K., Counts, S., Stecher, K., Hoff, A., Johns, P.: Pathfinder: An Online Collaboration Environment for Citizen Scientists. In: Proc. CHI 2009, pp. 239–248. ACM, New York (2009)
21. Minkler, M., Wallerstein, N. (eds.): Community Based Participatory Research for Health. Jossey-Bass, San Francisco (2003)
22. Mintz, D. (comp.): Technical Assistance Document for the Reporting of Daily Air Quality - the Air Quality Index (AQI). Tech. Research Triangle Park, U.S. Environmental Protection Agency (2009)
23. Modest Maps, <http://modestmaps.com>
24. National Audubon Society. History & Objectives. Christmas Bird Count, <http://www.audubon.org/bird/cbc/history.html>
25. O'Rourke, D., Macey, G.: Community Environmental Policing: Assessing New Strategies of Public Participation in Environmental Regulation. *Journal of Policy Analysis and Management* 22(3), 383–414 (2003)
26. Paulos, E., Honicky, R.J., Hooker, B.: Citizen Science: Enabling Participatory Urbanism. In: Foth, M. (ed.) *Handbook of Research on Urban Informatics*, pp. 414–436. IGI Global, Hershey (2008)
27. Tableau Server, <http://www.tableausoftware.com>
28. TIBCO, Spotfire Decision Site, <http://spotfire.tibco.com>
29. Tufte, E.: *Envisioning Information*. Graphics Press, Cheshire (1990)
30. Viégas, F.B., Wattenberg, M., van Ham, F., Kriss, J., McKeon, M.: ManyEyes: a site for Visualization at Internet Scale. *IEEE Transactions on Visualization and Computer Graphics* 13(6), 1121–1128 (2007)

Active Capacitive Sensing: Exploring a New Wearable Sensing Modality for Activity Recognition

Jingyuan Cheng¹, Oliver Amft², and Paul Lukowicz¹

¹ Embedded Systems Lab, University of Passau

{jingyuan.cheng,paul.lukowicz}@uni-passau.de

² Signal Processing Systems, TU Eindhoven

amft@tue.nl

Abstract. The paper describes the concept, implementation, and evaluation of a new on-body capacitive sensing approach to derive activity related information. Using conductive textile based electrodes that are easy to integrate in garments, we measure changes in capacitance inside the human body. Such changes are related to motions and shape changes of muscle, skin, and other tissue, which can in turn be related to a broad range of activities and physiological parameters. We describe the physical principle, the analog hardware needed to acquire and pre-process the signal, and example signals from different body locations and actions. We perform quantitative evaluations of the recognition accuracy, focused on the specific example of collar-integrated electrodes and actions, such as chewing, swallowing, speaking, sighing (taking a deep breath), as well as different head motions and positions.

1 Introduction

On-body sensing and activity recognition are a key concept in Pervasive Computing [1][6]. It enables a broad range of applications, from new mobile user interfaces, sports assistant systems, to health and assisted living. Today, the majority of on body activity recognition systems rely on motion sensors, such as accelerometers, gyroscope, magnetic field sensors, and combinations thereof [9][18][28]. On one hand it is due to the availability of cheap, miniaturised devices. On the other hand, motions of body parts are the key factor in almost all human activities. Despite their success motion sensors have some limitations:

1. Not all activities can be sensed from motion. For example, dietary monitoring has recently received significant interest in activity recognition. However, neither chewing nor swallowing can be easily detected by motion sensors [7].
2. Attaching motion sensors is not practicable for every body location. This is particularly true for hands and the head.
3. Signals from motion sensors can be ambivalent (as different actions are for example associated with similar motions) and noise (e.g. as sensor positions

shift). Thus, even if a particular activity can be captured using motion sensors, the accuracy could benefit from additional sensors that provide complimentary information and have different, independent sources of error.

As a consequence there has been significant interest in alternative sensing modalities. Textile stretch [14] and fibre optical sensors [13] have been proposed to detect posture. Another approach has been to add sensors for environmental parameters such as ambient sound, temperature, or air pressure [20][23].

Finally there are some more experimental approaches involving signals from “inside” the body. Body sound from the wrist has been used to detect hand motions [2] and from the ear to detect chewing [3]. In [10] the use of Electrooculographic eye tracking has been demonstrated for the recognition of reading activity. It has also been shown how to use force sensitive resistors to detect muscle [22][4] activity. Another interesting example is the use of radar directed at the body can detect vital signs [25].

The work presented in this paper falls into the above category of novel sensing approaches that attempt to utilise information from inside the body. It adapts the physical principle of capacitive sensing used in industry (for example to inspect closed boxes on a conveyor belt) to wearable activity sensing. In simple words, we consider a capacitor build out of a conductive textile electrode and the human body as dielectric. We then analyse capacitance changes caused by muscle motion, tissue displacement, electrode deformation, etc. This approach is attractive for activity sensing for the following reasons:

1. It provides information that is difficult to obtain with other unobtrusive sensors. For example, in the quantitative study presented in this paper, we use textile electrodes integrated in a collar to recognise among others, chewing and swallowing.
2. A sensor at a single location provides signals from a broad range of actions and physiological parameters. Thus, in addition to chewing and swallowing, we demonstrate the recognition of speaking, head motions (shaking, nodding), head positions, and deep breathing using the collar setup.
3. The sensing principle can be applied to different body locations. Besides collar and wrist setups, we show signals from the upper leg that can be used for modes of locomotion recognition and signals from the chest that are relevant for vital signs monitoring.
4. The system is based on textile electrodes that can be easily and unobtrusively integrated into clothing. It requires neither direct skin contact nor special fixation beyond the pressure of normal close fitting garments.

Related Work. Specifically for capacitive sensing previous work proposed using basically the same method to measure pulse on the wrist [26]. We have done a preliminary evaluation of this approach for pulse and breathing rate measurements on the chest [12]. There also exists a large body of work on capacitive coupling electrodes for heart rate monitoring and ECG, e.g. [21][27]. However, our work is based on a fundamentally different principle. Whereas our approach generates an electric field and measures the influence of capacitance changes due

to structural changes inside the body, the capacitive coupling electrodes cited above, measure the electric field generated by the body.

Capacitive sensing is widely used in industry for proximity sensing but moreover, to examine the content of closed boxes on a conveyor belt. Taking the idea further, there has been a significant amount of research on electric capacitance volume tomography [29] that attempts to reconstruct complex structures from multiple capacitive measurements.

The use of on-body capacitive sensing for user interfaces has been proposed by [31]. Later the same group has used on-body capacitive sensing for motion tracking in a dance application [8]. In this work the capacitive measurement had been used to measure distance between body parts, which is different from our approach. Capacitive gesture recognition for pervasive computing (but not wearable systems) has been discussed in [30] and in a string of other publication by this group. In the wearable field capacitive sensing is the basis of widely used textile pressure sensors as well [21]. Moreover, it was used for tracking people using electrode arrays embedded in a carpet [19] and as insole system measuring weight bearing [17]. To our knowledge, capacitive sensing had not been investigated for monitoring activities in the breadth attempted in this work.

Paper Contributions. We propose and evaluate a new way to derive activity-related information by “looking inside” the human body with a capacitance sensor.

Specifically the paper makes the following contributions:

- While capacitive sensing in itself is an established principle we have put forward a novel concept for using it in wearable activity recognition.
- Starting from extensive simulations, we designed and implemented the sensing hardware needed to deal with the specific requirements of our approach, in particular, the large dynamic range and very low electronics noise.
- We performed extensive experiments with different electrodes locations and activities. We present selected signals from those experiments and use them to explain the properties and potential of the proposed sensing approach.
- We performed quantitative evaluations of the recognition performance achieved with our system in the specific example of collar electrodes. Our recordings include the activities while working at a computer and while walking (to investigate the impact of motion artifacts).
- To further underscore the potential of this sensing approach, we present initial quantitative results (from the same collar electrode positions) for spotting swallowing in the continuous data stream, distinguishing between different swallowing amounts, and estimating respiration rate for shallow, normal, deep breathing.

We would like to point out that the aim of this work is **not** to prove the utility of the new modality for a particular real life application. Instead, we aim to establish a basic understanding of how to implement and use the modality and what sort of information it can provide. For a new sensing modality, such basic understanding is a necessary pre-condition for conceiving and demonstrating

concrete applications. Thus, we aim to lay the groundwork for other researchers to build on, when including this new sensing modality in their systems and enriching their future applications.

2 Sensing Principle

A capacitor is, in essence, a device that can store energy in an electric field. The best known example is the parallel plate capacitor, having two rectangular conductive plates separated by a gap filled with a non-conductive dielectric material. There are no specific requirements on the material from which the conductive plates are made. Thus, enabling conductive textile to be used, which means that they are very unobtrusive and easily integrated in clothing.

The electric field of a capacitor depends on the material placed between the plates, which can “dampen” the field. The damping depends on the molecular properties of the material as well as on its structure and shape. It can be as simple linear dependence like in the parallel plate capacitor, but also arbitrarily complex relationship reflecting elaborate shapes (including cavities) and inhomogeneities in the molecular properties.

The factor determining the voltage V that a given charge Q produces (distance between the plates and the influence of the material between them) are summarised as the capacitance C of the device. The key equation at the heart of the active capacitive sensing is $C = \frac{Q}{V}$, where Q and V can be easily measured, and C depends on the properties of an object including what is hidden inside it. Thus, object of changing structure, cavities, or changing surface, e.g. making it wet, will change C .

Hence, by measuring two electric parameters we can “look” inside an object in a non-invasive way. Clearly the information that we get is very limited: a single scalar value. Consequently, a capacitive measurement with a single electrode pair can not reveal complex structural information. However, it can indicate structural changes that take place within objects.

Capacitive Measurements on the Human Body. Figure 1 shows the simulated electric field distribution in the human body during a capacitive measurement. This data is part of extensive simulation that we have performed as part of our system design. We used the SEMCAD package [1]. The simulation was performed with the following configuration:

- Instead of using two electrode plates, we use just one. The second electrode is then effectively “earth”. This is a common approach in many capacitive systems (e.g. touchpads) and allows us to more easily integrate several close-by electrodes.
- We use AC current instead of DC to charge the capacitor. Since the capacitance of different materials varies with the oscillation frequency, this allows us to better optimise sensitivity of the system to certain effects.

Electric field intensity just below the electrode is several orders of magnitude higher than just a few cm further inside the body. Further inside the body the

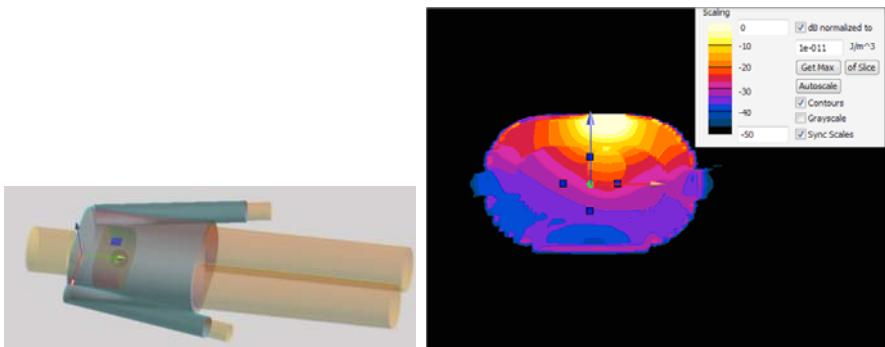


Fig. 1. Simulation of the electric field generate by an electrode on the chest. Left: the simulation setup. Right: the resulting field. Note that the different tones correspond to a dB (logarithmic) scale.

intensity is even lower. With regard to information that the capacitive signal could provide, the following conclusions can be drawn:

1. Any action that produces changes in the position of the electrode (in particular its distance to the skin) will have a very strong effect on the signal. This is on one hand a major source of noise, on the other, it can contain useful information related to motion or posture (when a person moves or changes posture the electrode will in general be displaced or deformed).
2. Any changes taking place directly below the electrode will produce a clear (although much weaker) signal. Such changes can be muscles flexing or hyoid movement during swallowing. The exact range of this regime varies depending on the setup between less than 1 cm to a few cm.
3. Changes deeper inside the body will only produce a distinguishable change in capacitance if they involve a large volume. A good example is breathing, where a large amount of air enters the lungs inside of the body.

3 Sensing Hardware

The overview of our sensing hardware is shown in the top part of Figure 2. We used four front-end boards to provide four independent channels, converting the capacitance into voltage. The voltage is AD-converted and sent out via ZigBee. We used a Tmote mini node for the wireless transmission to a Tmote-sky, which is connected to PC USB port.

The electrode itself is made of conductive textile, which is both thin and flexible. It can be easily integrated by cutting a required shape and sewing into the middle of 4 layers of soft paper (ink eraser tissues).

Front-end Boards. Our simulations have shown that existing commercial solutions for capacitive measurement are not sufficient to meet our demands:

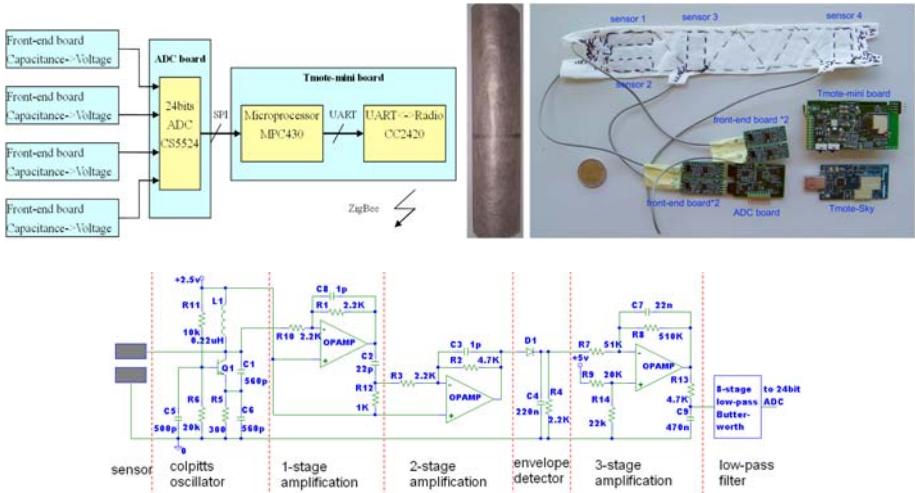


Fig. 2. Top left: top level diagram of the sensing hardware. Top right: implemented hardware including the electrode used for the neck experiments (left side). Bottom: schematic of the analog part of the front-end board.

- Small capacitance and ultra-low noise: the capacitance to be measured varies from several pF to several hundred pF, where information is the signal change, which could be as small as 0.01 pF.
- High measuring frequency: as confirmed by our simulation, the higher frequencies can better penetrate the body, and can thus provide more information inner-body changes.

These requirements have to be achieved in a small form factor, battery powered device. We have chosen to design and implement our own measurement circuit as shown on the bottom of Figure 2. The circuit is based on concept from [26]. The capacitor consisting of the conductive textile electrode, the human body, and ground is part of a copitts oscillator that generates a sinusoidal voltage. The oscillation frequency of the circuit is given by

$$f = \frac{1}{2\pi\sqrt{L(C_{circuit} + C_{sensor})}}, \quad (1)$$

where L and $C_{circuit}$ are the characteristic inductance and capacitance of the circuit, and C_{sensor} is the measured capacitance of the electrode. In our system $C_{circuit}$ is 17 MHz.

This sinusoidal signal from the copitts oscillator is differentiated by the capacitor and resistor after a 1st-stage amplification, converting the change of frequency to the change of amplitude. After a 2nd-stage amplification, which isolates and provides enough driving current, this change of amplitude (up to at most 100 Hz) is extracted by an envelope detector. The 3rd-stage amplification then amplifies the change into a proper input range of the ADC.



Fig. 3. Different sensor setups with which experiments were performed. Chest placement, wrist placement, and neck placements. For the neck setup, sensor placement in an elastic band and integration in a pullover collar are shown.

Because the distance between sensor and skin affects the results most, the circuit must provide both broad measuring range and high precision. A 24-bit ADC was used for this purpose. At the same time, signal noise must be suppressed. Common noise removing methods as isolating the power supply with appropriate capacitors or adding resistor for impedance matching, do not work in our case, because they are meant to remove high frequency noise only. Thus, we optimised our hardware design by separating digital and analog circuits, amplifying and digitising close to the front-end, and using multi-stage low-pass filters. Because we focused on human body activity, sample rate was set to 40 Hz, with low-pass filters' 3dB frequency fixed to ~ 15 Hz. In addition, we used ultra-low noise DC-DC voltage regulators to provide amplifier reference voltage.

For multi-channel measurements, oscillation frequencies of the individual channels must be distinctive. We chose L1 to 0.33 μ H, 0.47 μ H, 0.68 μ H, and R3 correspondingly to 470Ω , 680Ω , and 680Ω . Further, the insulation material between sensor and skin should be either of a high dielectric coefficient or thick enough to avoid crosstalk on nearby channels.

4 Signals Analysis

Using the hardware described in the previous section we have performed extensive experiments with different placement of electrodes and activities to understand what type of signal the proposed method can provide. In this section we give some interesting examples, that illustrate the points made above and relate to the systematic evaluation that are described in Section 5.

Chest Electrodes. In Figure 4 we first look at the signal collected by electrodes on the chest (see 3), showing breathing cycles. This is an example of the third category described above: signals originating far from the electrode, but involving large body volume. The signal from the left electrode has a superimposed component related to the heart beat, as the electrode is in close proximity to the heart. Heart beat becomes more visible when the wearer holds his breath.

Wrist Electrodes. The properties of the proposed sensing approach are well illustrated by the signal from a wrist electrode which is shown in Figure 4. Although the electrode is by far not near to the lungs, there is a clear breathing signal. Even though the electrode is placed far from the chest, the electric field

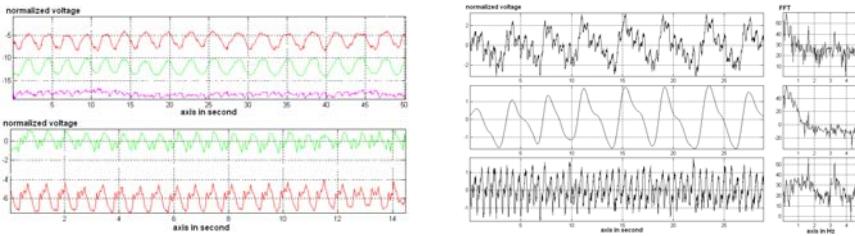


Fig. 4. Left top: breathing and pulse signals from chest electrodes. Signals from left and right electrodes (top and middle trace) and their difference (bottom trace). Left bottom: signals from front and back chest electrode at higher amplification, when wearer is holding breath. The pulse can be clearly seen then. Right: signal from a wrist mounted electrode and spectrogram. Right top: signal showing a mixture of pulse and breathing. Right middle: low pass filtered signal showing breathing. Right bottom: high pass filtered signal showing pulse.

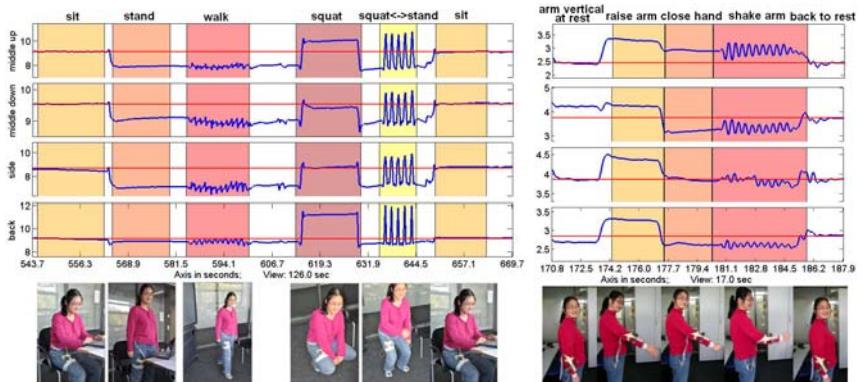


Fig. 5. Left: signals from upper leg electrodes (front middle front up, side and back) during a modes of locomotion experiment. Right: signals from lower arm and wrist electrodes during a movement sequence.

still passes through and around the chest when going towards “earth”. This means that the large volume breathing effect is still visible.

The pulse signal is clearly seen even when breathing. In fact, it is even clearer than from the electrode above the heart. This is because the wrist contains many blood vessels directly below the skin surface and the distance to the electrode matters much more than volume of change. This was illustrated in Figure 1.

As Figure 5 shows, arm and wrist signals also contain activity-related information. Raising the arm, closing the hand, or shaking the arm, all produce distinct patterns related to muscle and tissue motion. Note that the muscles and tendons moving fingers are extending from the lower arm, which is why the hand closing gesture (and potentially other palm and finger motions) can be distinguished at the wrist.

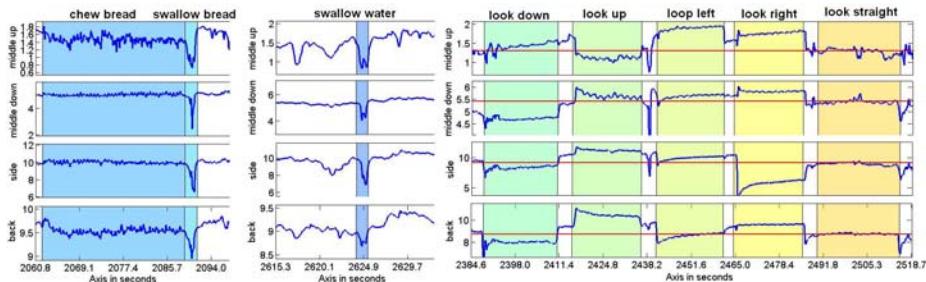


Fig. 6. Signals from the neck electrodes. Left: chewing a piece of bread and swallowing. Middle: swallowing 15 ml of water. Right: different head positions.

Upper Leg Electrodes. The recognition of modes of locomotion (differentiation between sitting, standing, walking etc.) is a standard problem in activity recognition. Figure 5 left shows how the sensor can be used for this purpose, with the electrodes wrapped around the upper leg. Walking, sitting standing and doing crunches all produce different signal patterns. They are mostly due to muscle shape changes compressing top level tissue and skin (potentially also the sensor material). An interesting questions for future research is whether careful electrode placement and elaborated signal processing can provide cues on activation and state of different muscles.

Neck Electrodes. We investigated electrode positions at the neck (see Figure 3). This position was chosen for three reasons.

1. It is a rich source of information as head motions, positions, speaking and chewing, all cause skin and muscle motions directly below the skin. The hyoid moves as people swallow. Veins are covered by thin tissue only.
2. Many of the activities to which those signals relate, are difficult to detect using other non-obtrusive sensors. Head position and motion requires head mounted sensors, which is not always practical. For chewing and swallowing most existing solutions require electrodes to be glued to the skin (although we have also used sound from the ear in previous work [3]).
3. People are used to wearing things like scarfs, ties, collars, etc. on the neck. Our electrodes are just pieces of textile and can be unobtrusively integrated.

Figure 6 shows the signals from chewing and swallowing. Chewing is best seen in front upper electrode as skin motion and deformation caused by jaw motions. To a much lesser extend the signal is present in the other three electrodes. The swallowing signal shape is very clear: in the front electrodes it has a “W”-like shape caused by the hyoid moving up and down. Swallowing water and swallowing bread causes different shapes.

Figure 6 illustrates our analysis of head postures (left, right up, down). Clear differences can be seen in the amplitude levels on the different electrodes. These are due to electrode deformations, tissue compression, and skin movement. The

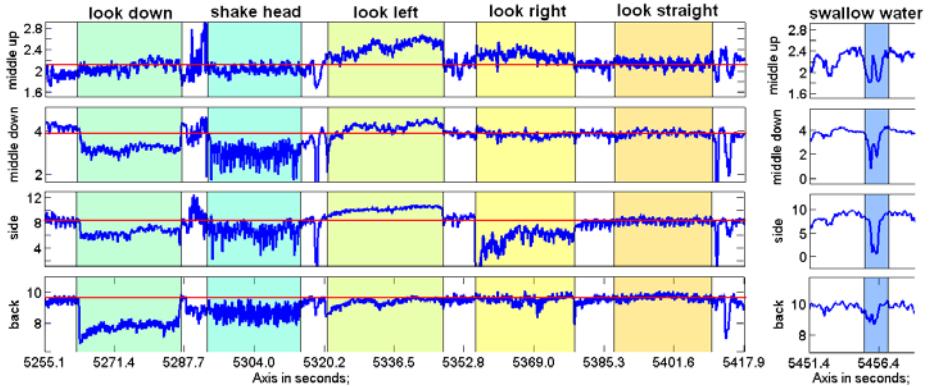


Fig. 7. Influence of motion artifacts on the neck electrodes. Left: signals for different head positions. Right: swallowing water while walking.

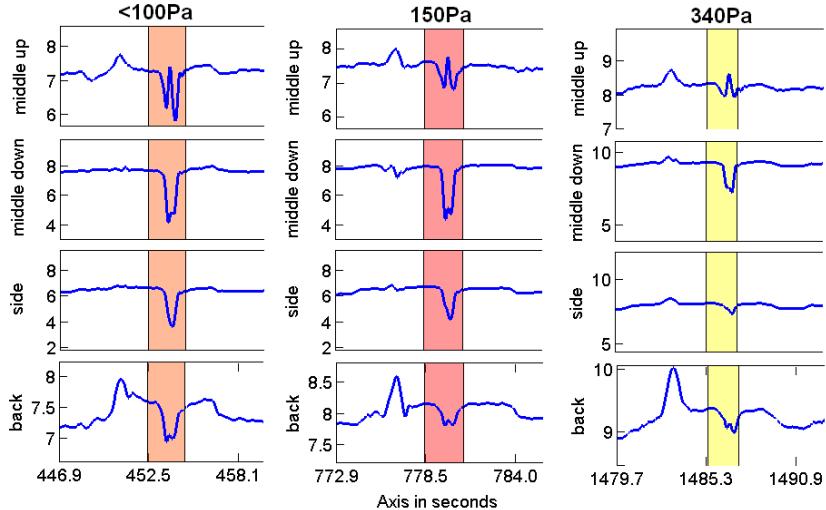


Fig. 8. Swallowing signal when the collar presses the electrode against the neck with 100 Pa, 150 Pa, and 340 Pa

same factors are responsible for the very articulate rhythmic signals for nodding and head shaking. Finally, speaking shows a strong but very variable signal. Nevertheless, in Section 5, we will show that it is distinct enough for a reasonable recognition results.

Motion Artifacts and Electrode Attachment. Previous paragraphs have detailed that the proposed sensor is highly sensitive to a broad range of factors. Thus, motion artifacts and sensor attachment are obvious concerns. To illustrate the effect of motion artifacts, signal for different head position and for swallowing, recorded while a person was walking, are shown in Figure 7. While

the noise due to walking can be clearly seen, key features of the specific activity remain visible. Overall, motion artifacts influence signal quality, however they do not completely obscure the signal information content. This is confirmed by dedicated recognition experiments in Section 5.

Concerning sensor attachment, a primary question is how tight the electrode must be pressed onto the body. During the recognition experiments participants were assisted to attach the collar “tight but comfortable”. For a more quantitative assessment, we have recorded sample signals while measuring the force, which the collar extended. From this force the pressure on the neck was estimated. The results for pressures of 100 Pa (band extended by 1 mm), 150 Pa and 340 Pa (for comparison, the atmospheric pressure is around 100'000 Pa) are shown in Figure 8. Interestingly, the least pressure leads to the best signal. This is because the main contribution to the signals comes from deformation of skin and soft tissue directly below the electrode, which is suppressed when the collar is too tight.

5 Quantitative Evaluation

In this quantitative evaluation we focus on the collar setup. As detailed in Section 4 this setup provides information related to different activities, such as head motions, chewing and swallowing, which are difficult to detect with other unobtrusive sensors. For the same reason it is also challenging, as the system needs to deal with a broad range of variable signals.

We proceeded in four stages. First we investigate how well the signals corresponding to 11 different activities can be differentiated in isolation. Thus, we check how much relevant information is contained in the signal. Secondly, we investigate how well the swallowing can be spotted in the continuous data stream. We choose swallowing because it is a short subtle signal (as opposed to activities, such as nodding which are longer and repetitive). Swallowing spotting is also relevant for nutrition related applications. Here we demonstrate that relevant activities are not “swamped” by noise from the NULL class. Thirdly, we attempt to distinguish different swallow sizes, this tests the limits of information we can extract. In further testing these limits, we attempt to distinguish different breathing modes, which can be relevant for many sports and health-related applications.

5.1 Recognition of Activities

Experimental procedure. Three subjects (one female, two male; aged between 25 and 45 years) have worn the electrode collar during computer work and when walking in corridors of the office building. In both scenes we asked them to perform a set of head movements for 20 s each: nodding, shaking head, look down, up, left, right, and straight. Moreover, the individuals were asked to drink and swallow water from a cup ($10 \times \sim 6 \text{ ml}$), chew and swallow bread pieces (total $5 \times 2 \text{ cm}^3$), and speak (reading a text aloud from the computer screen or talking

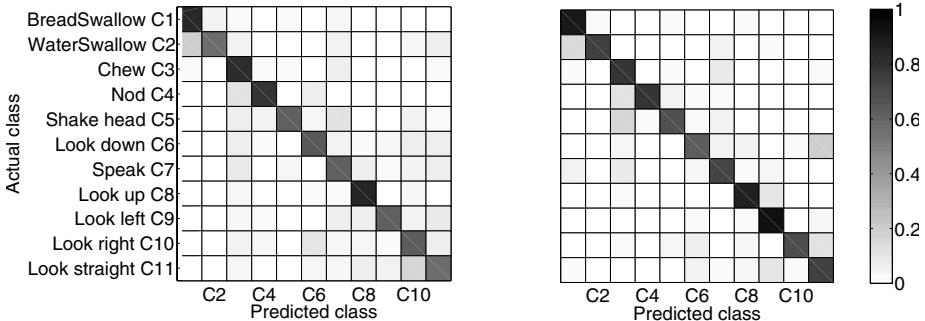


Fig. 9. Activity recognition confusion matrices using three capacitive sensors embedded in the collar. Left: activities while sitting and walking, Acc = 0.69. Right: activities while sitting, Acc = 0.77.

to the experiment observer for 20 s). All actions were repeated for 3 times, in order to introduce natural variability in the recordings. “Nodding” and “looking up” were recorded for the sitting scene only, as these were both exhausting for the participants, and safety critical during the corridor walking. All recordings for computer work and walking were made in one session for each participant.

To maintain electrode position the collar was fixed with an elastic band, the subjects were told to fix it so that it is tight but comfortable. We analysed two front, one side, and one back electrode position in the same way as presented for our signal study in Section 4 above. Preliminary analysis showed however, that the back position did not provide useful information. Thus we did not consider it here. An experiment observer controlled the recording and advised participants on the activities to perform. In addition all recordings were captured on video. The recording lasted for about 70 minutes for each individual, in total ~ 4.3 hours of data were acquired. The observer annotated all activities during the recording. These annotations were refined in post processing step based on signal waveforms.

A particular challenge is to accurately annotate natural swallowing [5]. Typically, the swallowing reflex is initiated unconsciously, which makes it difficult to identify and annotate it during recordings. In this study, we asked the participants to indicate swallowing with a hand sign. In addition, we reviewed the sensor data and video material to decide unclear cases of potential swallows. For this purpose we installed the video camera in the computer work scene such that it captured the hyoid movement. This procedure is a standard technique used in swallowing analysis [5][15]. For the walking scene, we had to rely on participant hand signs and review of sensor waveforms.

Analysis method and results. The analysis employed a linear discriminant classifier. Time domain features, such as signal mean, variance, maximum, etc. were derived from all three sensors (45 in total) in sliding windows of 1.5 s length without overlap. We employed 10-fold cross-validation to obtain training (9 parts) and

testing (1 part) observations. The training was controlled to avoid class skew. The classification output was compared to our annotation and the class-normalised accuracy computed. We first performed a combined analysis over the sitting and walking segments. For comparison, we analysed the sitting activities separately as well. “Nodding”, “looking up”, and “chewing bread” were excluded from the walking scene for practicability reasons.

Our combined results for sitting and walking activities showed an accuracy of 69%. For the sitting activities 77% were achieved. Figure 9 shows the classifier confusion matrices for both analyses. While the different head postures and movements incurred some confusions, we observed that particular activities such as speaking and chewing could be very well discriminated. Moreover, the discrimination of fluid and bread swallowing is remarkable. Although including walking reduces performance, these results indicate that the sensor can provide useful information in the presence of motion artifacts.

5.2 Spotting of Swallowing

We analysed the viability of spotting swallowing events in the continuous sensor data, as it is an essential component of food and fluid intake [7]. In particular, we were interested to analyse the spotting performance with regard to swallowing pattern variability and the effect of artifacts, such as walking. For this analysis we focused on fluid swallows and utilised the same experimental data as presented in Section 5.1 above. As this dataset included a variety of other activities, we could test the spotting performance under realistic conditions.

Analysis method and results. We used an online pattern spotting procedure, Feature Similarity Search (FSS), developed in previous work [6] to evaluate swallowing detection performance in this work. The procedure uses trained feature patterns to continuously search the sensor data. In the search step, a variable observation window is used to derive features from data section. Between these features and the trained feature pattern we computed the Euclidean distance to compare and select a section. For the selection, a threshold was applied on the computed distances. At each time point only one such section can be correct, hence potential sections are kept in a buffer, until no further section could overlap with such already ones. Both, selection threshold and the variable window bounds were derived during the training step.

In this analysis we used time domain features from all three capacitive sensors, including the same types as described in Section 5.1 above, and three additional feature sets of the same type, describing three equally sized partitions of the section under investigation. This approach allows to convert the temporal swallowing signal pattern into a spatial one. The search was performed at constant time intervals of 0.25 s. We used a 10-fold cross-validation by splitting the dataset into 10 partitions and using nine for training and one for testing at each iteration. The partitions were controlled to not intersect with the swallowing sections.

In our evaluation, we included sitting and walking scenes to study spotting performance under noisy pattern condition. For comparison, we also analysed

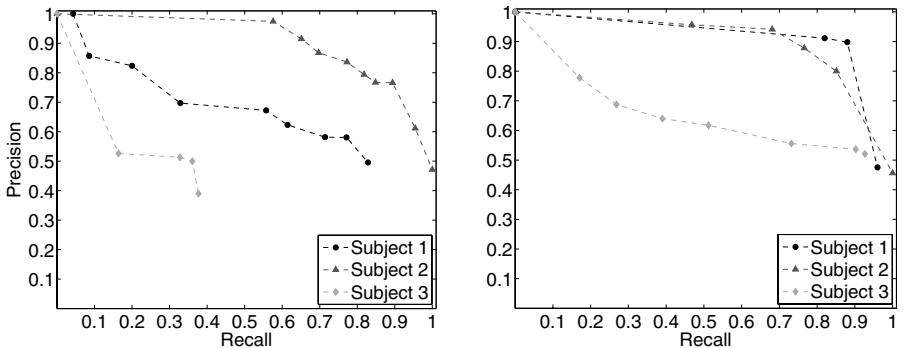


Fig. 10. Precision-recall tradeoff for spotting swallowing in three subjects. Left: swallowing while sitting and walking. Right: swallowing while sitting.

the sitting scene separately. In total 212 swallows were recorded, amounting to 2.8% of the dataset size. Figure 10 shows the precision-recall tradeoff for both spotting situations and the three participants. We observed that the performance was more variable in the combined sitting and walking analysis. Performance increased for the sitting scene, as to be expected from the reduced motion noise. The results show that the spotting is feasible, in particular when not walking. Under calm conditions a performance of 80% recall at 60% precision or more, can be expected. Although our dataset is smaller than the ones investigated for acoustic and electromyography swallowing spotting in earlier works [5], these results are very promising.

5.3 Swallowing Amount Estimation

We investigated the classification of drinking sizes as this information could be used to estimate fluid consumption. For this investigation we used the same collar and setup as in the activity recognition analysis.

We asked three individuals (one female, two male; aged between 25 and 30) that were not the same as in the activity recognition analysis to drink 5 ml and 15 ml water amounts. The amount was controlled using a calibrated glass. The experiment observer filled the glass for each drink. The participants were asked to swallow the fluid at once. A sequence of 10 swallows of each amount was taken and the sequence was repeated for three times, resulting in ~ 30 swallows per amount. The recording was annotated and post-processed as the ones before.

Swallowing amount recognition was performed using the features variance, minimum, and maximum from all three sensors (9 features total). The linear discriminant classification was used. Figure 11 shows the Receiver Operator Characteristic (ROC) performance analysis obtained from the classification result. From the ROC, we computed the “Area under the curve” (AUC) for quantitative performance estimation. The results show that a similar AUC can be achieved for all participants. ROC and AUC are the most appropriate illustrations for this two-class problem. As Figure 11 illustrates, is the performance

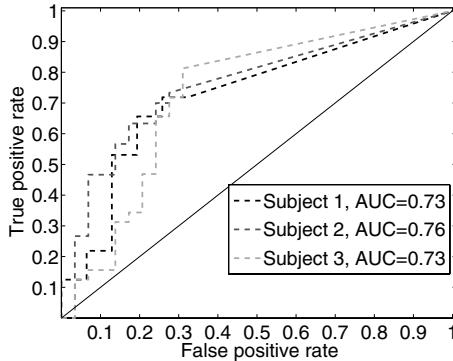


Fig. 11. ROC analysis for classifying 5 ml and 15 ml water swallowing. AUC values indicate the area under the curve.

clearly above the level of random choice. This result indicates that the sensors could be used to assess fluid amount. The result is in a comparable range to previous investigations using audio and Electromyography [5].

5.4 Respiration Rate Detection

As our preliminary signal study showed that breathing could be observed in the sensor data, we investigated the respiration rate detection from the capacitive collar system. We studied breathing with the same individuals who participated in the swallowing analysis (Section 5.3 above).

We asked the participants to breathe in and out in three qualitative modes: deep, normal, and light. We recorded 10 breathing cycles during walking and standing of each mode and repeated this protocol three times. In total ~ 30 breathing cycles were recorded per mode, participant, and scene. As not all participants achieved exactly 30 cycles, the numbers were noted and checked in the waveforms. Our post-recording analysis showed that the breathing was difficult to identify from the waveforms during walking, hence we could not verify the number of breaths. Henceforth, we concentrated our analysis on the standing scene.

For this study we chose the capacitive sensor at the neck side, as this showed the largest amplitudes during breathing for all participants. The signal was band-pass filtered using a fourth-order Butterworth filter with $f_{Low} = 0.3$ Hz and $f_{High} = 2$ Hz. These frequency ranges reflect the natural variation of the respiratory rate in adults. On the resulting signal a hill-climbing peak detection algorithm was applied with thresholds for positive and negative slope set to $\frac{g}{2}$ of the considered signal. To set the thresholds automatically, a longer observation period could be used, which contains several breathing cycles at high probability. The resulting peak detection count were compared to the annotated counts and the accuracy was computed. Figure 12 shows the detection performance for all participants and breathing modes.

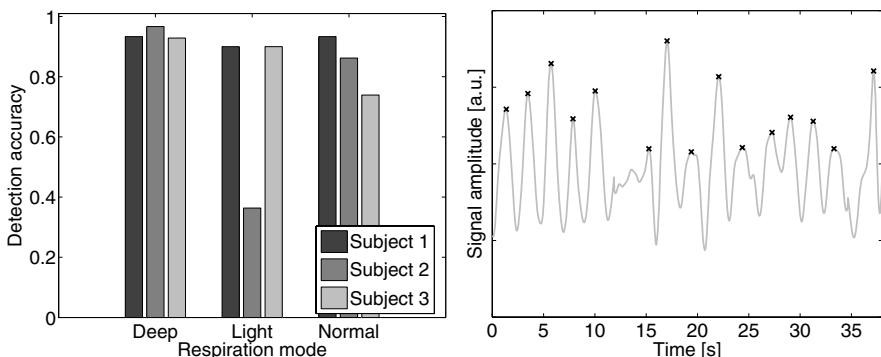


Fig. 12. Respiration rate detection results. Left: accuracy for all participants and breathing modes. Right: deep breathing detection sample.

Although our algorithm performed well to identify breathing, it might fail, if the breathing is held. In this case the peak detection algorithm may pick the heart rate, which is in a similar frequency band and amplitude level as the light breathing. However, under normal breathing using the electrode position at the neck, heart beat is marginally disturbing the breathing detection.

6 Conclusion

While being at an early state, the proposed application of capacitive sensing shows promising results, making it appealing for a wide range of applications. Starting from electric field, we demonstrated that this new sensing concept is well suited for retrieving activity and physiology-related information at multiple body locations. This is particularly interesting as our capacitive sensors are based on textile patches and can be conveniently integrated into regular clothing.

Since head-related activities are key to many activity recognition applications, we selected a collar system for our quantitative analysis. We observed that our approach is sensitive to motion, body shape, and tissue changes in a spectrum of activities, while providing useful information even under noisy walking conditions. From these results, we concluded that capacitive sensing is a viable and highly interesting new sensing concept for wearable monitoring. Moreover, from analysing swallowing and breathing we have seen that the sensor has particular features that promote its consideration for biomedical investigations and healthcare.

In summary, we expect that capacitive sensing will have a vital prospect as modality that is complementary to established concepts in activity monitoring. This work has initially demonstrated its potential. Further work should address the integration and optimisation for individual applications, which can even increase the sensor's reliability.

References

1. Abowd, D., Dey, A., Orr, R., Brotherton, J.: Context-awareness in wearable and ubiquitous computing. *Virtual Reality* 3(3), 200–211 (1998)
2. Amento, B., Hill, W., Terveen, L.: The sound of one hand: a wrist-mounted bio-acoustic fingertip gesture interface. In: Conference on Human Factors in Computing Systems, pp. 724–725. ACM, New York (2002)
3. Amft, O., Stager, M., Lukowicz, P., Troster, G.: Analysis of Chewing Sounds for Dietary Monitoring. In: Beigl, M., Intille, S.S., Rekimoto, J., Tokuda, H. (eds.) UbiComp 2005. LNCS, vol. 3660, pp. 56–72. Springer, Heidelberg (2005)
4. Amft, O., Junker, H., Lukowicz, P., Tröster, G., Schuster, C.: Sensing muscle activities with body-worn sensors. In: BSN 2006: Proceedings of the International Workshop on Wearable and Implantable Body Sensor Networks, pp. 138–141. IEEE Press, Los Alamitos (2006)
5. Amft, O., Tröster, G.: Methods for detection and classification of normal swallowing from muscle activation and sound. In: PHC 2006: Proceedings of the First International Conference on Pervasive Computing Technologies for Healthcare, ICST, pp. 1–10. IEEE digital library, Los Alamitos (2006)
6. Amft, O., Tröster, G.: Recognition of dietary activity events using on-body sensors. *Artificial Intelligence in Medicine* 42(2), 121–136 (2008)
7. Amft, O., Tröster, G.: On-body sensing solutions for automatic dietary monitoring. *IEEE Pervasive Computing* 8(2), 62–70 (2009)
8. Aylward, R., Paradiso, J.: Sensemble: a wireless, compact, multi-user sensor system for interactive dance. In: Proceedings of the 2006 conference on New interfaces for musical expression, pp. 134–139. IRCAM—Centre Pompidou Paris, France (2006)
9. Bao, L., Intille, S.: Activity recognition from user-annotated acceleration data. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 1–17. Springer, Heidelberg (2004)
10. Bulling, A., Ward, J., Gellersen, H., Troster, G.: Robust Recognition of Reading Activity in Transit Using Wearable Electrooculography. In: Indulska, J., Patterson, D.J., Rodden, T., Ott, M. (eds.) PERVASIVE 2008. LNCS, vol. 5013, pp. 19–37. Springer, Heidelberg (2008)
11. Chavannes, N., Tay, R., Nikoloski, N., Kuster, N.: Suitability of FDTD-based TCAD tools RF design of mobile phones. *IEEE Antennas and Propagation Magazine* 45(6), 52–66 (2003)
12. Cheng, J., Lukowicz, P.: Towards wearable capacitive sensing of physiological parameters. In: Second International Conference on Pervasive Computing Technologies for Healthcare, Pervasive Health 2008, pp. 272–273 (2008)
13. Dunne, L., Walsh, P., Smyth, B., Caulfield, B.: Design and evaluation of a wearable optical sensor for monitoring seated spinal posture. In: 2006 10th IEEE International Symposium on Wearable Computers, pp. 65–68 (2006)
14. Farrington, J., Moore, A., Tilbury, N., Church, J., Biemond, P.: Wearable Sensor Badge & Sensor Jacket for Context Awareness. In: Proceedings of the Third International Symposium on Wearable Computers, pp. 107–113 (1999)
15. Firmin, H., Reilly, S., Fourcin, A.: Non-invasive monitoring of reflexive swallowing. In: Speech, Hearing and Language: work in progress, vol. 10, Department of Phonetics and Linguistics. University College London, UCL (1997)
16. Gellersen, H., Schmidt, A., Beigl, M.: Multi-Sensor Context-Awareness in Mobile Devices and Smart Artifacts. *Mobile Networks and Applications* 7(5), 341–351 (2002)

17. Hurkmans, H.L.P., Bussmann, J.B.J., Benda, E., Verhaar, J.A.N., Stam, H.J.: Accuracy and repeatability of the pedar mobile system in long-term vertical force measurements. *Gait & Posture* 23(1), 118–125 (2006)
18. Kern, N., Schiele, B., Schmidt, A.: Multi-sensor Activity Context Detection for Wearable Computing. In: Aarts, E., Collier, R.W., van Loenen, E., de Ruyter, B. (eds.) *EUSA 2003. LNCS*, vol. 2875, pp. 220–232. Springer, Heidelberg (2003)
19. Lauterbach, C., Glaser, R., Savio, D., Schnell, M., Weber, W., Kornely, S., Stb'hr, A.: A Self-organizing and Fault-Tolerant Wired Peer-to-Peer Sensor Network for Textile Applications. In: Brueckner, S.A., Di Marzo Serugendo, G., Karageorgos, A., Nagpal, R. (eds.) *ESOA 2005. LNCS (LNAI)*, vol. 3464, pp. 256–266. Springer, Heidelberg (2005)
20. Lester, J., Choudhury, T., Borriello, G.: A Practical Approach to Recognizing Physical Activities. In: Fishkin, K.P., Schiele, B., Nixon, P., Quigley, A. (eds.) *PERVASIVE 2006. LNCS*, vol. 3968, pp. 1–16. Springer, Heidelberg (2006)
21. Lopez, A., Richardson, P.: Capacitive electrocardiographic and bioelectric electrodes. *IEEE Transactions on Biomedical Engineering*, 99 (1969)
22. Lukowicz, P., Hanser, F., Szubski, C., Schobersberger, W.: Detecting and Interpreting Muscle Activity with Wearable Force Sensors. In: Fishkin, K.P., Schiele, B., Nixon, P., Quigley, A. (eds.) *PERVASIVE 2006. LNCS*, vol. 3968, pp. 101–116. Springer, Heidelberg (2006)
23. Maurer, U., Rowe, A., Smailagic, A., Siewiorek, D.: eWatch: A Wearable Sensor and Notification Platform. In: Proceedings of the Int. Workshop on Wearable and Implantable Body Sensor Networks (2006)
24. Meyer, J., Lukowicz, P., Troster, G.: Textile Pressure Sensor for Muscle Activity and Motion Detection. In: Proceedings of the 10th IEEE International Symposium on Wearable Computers, pp. 69–72 (2006)
25. Michaelles, F., Wicki, R., Schiele, B.: Less Contact: Heart-rate detection without even touching the user. In: Proc. of the Eighth International Symposium on Wearable Computers, pp. 4–7. IEEE Computer Society, Washington (2004)
26. Oum, J., Lee, S., Kim, D., Hong, S.: Non-contact heartbeat and respiration detector using capacitive sensor with Colpitts oscillator. *Electr. Letters* 44(2), 87–88 (2008)
27. Ueno, A., Akabane, Y., Kato, T., Hoshino, H., Kataoka, S., Ishiyama, Y.: Capacitive sensing of electrocardiographic potential through cloth from the dorsal surface of the body in a supine position: a preliminary study. *IEEE Transactions on Biomedical Engineering* 54(4), 759–766 (2007)
28. Van Laerhoven, K., Cakmakci, O.: What shall we teach our pants? In: The Fourth International Symposium on Wearable Computers, pp. 77–83 (2000)
29. Warsito, W., Marashdeh, Q., Fan, L.: Electrical Capacitance Volume Tomography. *IEEE Sensors Journal* 7(4), 525 (2007)
30. Wimmer, R., Holleis, P., Kranz, M., Schmidt, A.: Thracker-Using Capacitive Sensing for Gesture Recognition. In: Proceedings of the 26th IEEE International ConferenceWorkshops on Distributed Computing Systems. IEEE Computer Society, Washington (2006)
31. Zimmerman, T.G., Smith, J.R., Paradiso, J.A., Allport, D., Gershenson, N.: Applying electric field sensing to human-computer interfaces. In: CHI 1995: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 280–287. ACM Press/Addison-Wesley Publishing Co. (1995)

Using Height Sensors for Biometric Identification in Multi-resident Homes

Vijay Srinivasan, John Stankovic, and Kamin Whitehouse

Department of Computer science, University of Virginia,

Charlottesville, Virginia, USA

{vs8h, stankovic, whitehouse}@cs.virginia.edu

Abstract. In this study, we evaluate the use of height for biometric identification of residents, by mounting ultrasonic distance sensors above the doorways in a home. Height sensors are cheap, are convenient for the residents, are simple to install in an existing home, and are perceived to be less invasive than cameras or microphones. Height is typically only a weak biometric, but we show that it is well suited for identifying among a few residents *in the home*, and can potentially be improved by using the history of height measurements at multiple doorways in a tracking approach. We evaluate this approach using 20 people in a controlled laboratory environment and by installing in 3 natural, home environments. We combine these results with public anthropometric data sets that contain the heights of residents in 2077 elderly multi-resident homes to conclude that height sensors could potentially achieve at least 95% identification accuracy in 95% of elderly homes in the US.

1 Introduction

The ability to identify residents in a home is crucial for many smart home applications: in order to respond to activities in the home, the system must be able to identify *who* is in a particular location or performing a particular action such as cooking or exercising. Existing innovative implementations that perform resident identification and tracking have several advantages, but also have drawbacks. Some approaches are inconvenient because they require the user to wear a tag [21,13], or to manually trigger a biometric sensor such as a thumbprint or retina scanner [18]. Some systems require cameras for gait, form, or face recognition [16], but cameras are often perceived as invasive because they can be used to collect much more information than just the user's identity [1]. Other implementations require structural changes to the home, such as instrumenting the floor [6,9] with force plates, which can incur high cost and effort. Many practical smart home applications such as in-home medical care for the elderly [17,2] and occupant-based energy monitoring [8] cannot use solutions that inconvenience the user, are intrusive, or require an expensive building retrofit. Our recent discussions with a commercial peace of mind elderly monitoring enterprise [4] reveal several interesting user requirements for accurate, long term elderly resident identification and tracking in homes: (1) residents will not wear tags or

manually identify themselves at every room for long periods of time, (2) residents will not allow perceived invasive devices such as cameras or microphones in the home, and (3) residents want the sensors to be fairly invisible, similar to existing motion sensor installations, and do not want an expensive building retrofit. Since existing implementations have some drawbacks with respect to the above requirements, commercial deployments by the elderly monitoring enterprise [4] today are limited to single-resident homes or do not fully monitor information about multiple residents.

This study examines the use of biometric height sensors to satisfy the above requirements for both the elderly monitoring enterprise [4], and a wide variety of other smart home applications. Height sensors have several advantages over existing approaches: they are cheap, convenient and minimally invasive for the residents, and not very time consuming to install in an existing home. We use ultrasonic distance sensors mounted above the doorways in a home to measure the height of individuals that walk through the doorway. The inherent accuracy of height sensing is too low for reliable biometric identification from a large population of individuals: it requires a **7cm** difference in height to differentiate people with 99% accuracy, and most people have heights within a small range from 160-180 cm. However, we make two key insights that allow height to be an effective biometric sensor *in the home*. First, most homes have very few residents: height may be a weak biometric for differentiating between 20 or more people, but is likely to be very effective in homes that have only 2-4 residents. Second, people move through a home in predictable ways, as determined by the floor layout: if height sensors are placed above every doorway, then the history of height measurements can be used to potentially surpass the inherent accuracy of the sensor.

The main contribution of this work is to demonstrate that height can be effective for biometric identification *in the home*. We evaluate the use of height as a biometric in four ways: (1) We quantify the biometric error of our approach using 20 subjects in a controlled laboratory environment, in which we vary the direction, speed, and location of the person walking under the doorway. (2) We measure the degree to which height sensors can identify room occupancy of residents in 3 natural home environments for 5 days each. (3) We use public anthropometric data containing the heights of elderly residents in 2077 multi-resident homes from the 2006 health and retirement study to estimate that our approach can potentially achieve at least 95% identification accuracy in **85%** of elderly US homes sampled in this study. (4) Through a simulation study, we show that incorporating the history of height measurements at multiple doorways using a tracking approach can potentially increase the proportion of homes where our solution is applicable with 95% accuracy from 85% to 95%, and also reduce the height difference required for 99% identification accuracy from 7cm to 3.25cm. We quantitatively compare our approach against two other state of the art non-invasive resident identification implementations, namely anonymous binary sensor and activity model based multi-resident tracking [20], and weight sensing [9], and find that our approach achieves improvements in identification

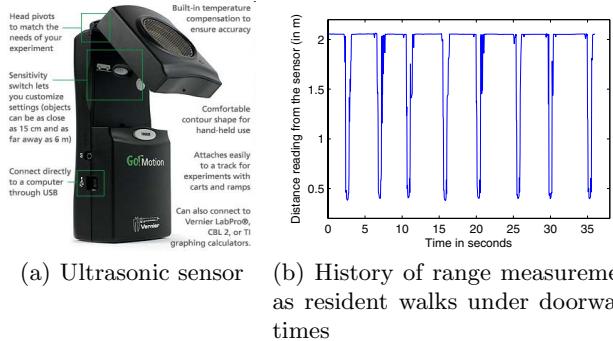


Fig. 1. Our study used the Go Motion ultrasonic range finder mounted above doorways (a). As users walked beneath the sensor, the range measurements changed (b).

accuracy or installation effort and cost compared to existing implementations of these approaches.

The rest of the paper is organized as follows. Section 2 discusses existing resident identification solutions from the literature. Section 3 gives an overview of our approach and describes our algorithm to sense height from ultrasonic distance measurements. Sections 4 and 5 describe the results of our controlled lab experiments and natural, in-home experiments respectively. Section 6 analyzes how our empirical results potentially extrapolate to a national level using public anthropometric data. Section 7 evaluates the potential improvement in the inherent identification accuracy of the height sensor by tracking the history of height measurements at multiple locations in a home. Section 8 discusses the application of height sensors for in-home room level tracking, and systematically lists the advantages and limitations of our current approach and study. We conclude by summarizing our findings in section 9.

2 Existing Solutions for Resident Identification

Resident identification in smart homes is a long-standing problem with many existing solutions. In this section, we discuss a representative sample of these solutions, including their advantages and disadvantages, and their applicability with respect to our user requirements.

Tag-and-track approaches operate by placing a uniquely identifiable device on each individual resident. This approach has been widely used since the Active Badge system almost two decades ago [9], and in other systems since. The pedestrian localization system proposed by Woodman et al [21] uses a foot mounted inertial sensor to track pedestrians to within .7 meters 95% of the time in a large office building with no additional infrastructure. More recently, innovative tracking solutions that require a very low infrastructure cost [2,15] are emerging. Tag and track approaches have three important advantages: (a) High location granularity with little or no infrastructure, (b) Selective preservation of

location privacy by switching off device, and (c) Highly scalable with respect to the number of residents in an indoor space: each user can be given a new device with a uniquely identifying number. However, one drawback of this approach is that it requires the user to actively carry the device at all times when location information is desired. It can be an inconvenience to in-home residents for long-term deployments. In our past experience with deployments, and while trying to use MoteTrack [13] wireless tags to collect ground truth in this study, users frequently forget to carry their tag, especially immediately after waking up or showering. Automatically reminding residents to carry the device is an option, but we believe such an approach is intrusive and inconveniences the user for long-term applications such as elderly medical monitoring.

Several indoor resident identification systems use **cameras** for computer-vision based face, shape, and gait recognition [16], or **microphones** and audio signal processing. These approaches might require expensive on-board computation or high communication bandwidth to a central base station that executes the vision algorithms, but are passive and highly accurate. However, user studies by researchers from Intel and companies like WellAware have found that a large fraction of potential users have perceived privacy concerns about cameras or microphone sensors [11]. Therefore, this class of approaches is most appropriate for short-term situations in which rapid deployment, and/or high accuracy are important, and where long-term privacy concerns of monitoring residents in their own homes are not an issue.

Wilson et al, in 2005, propose using only resident **usage models of anonymous motion sensors in rooms and switch sensors on daily-use objects**, and **resident activity models** to identify and track their activities and locations [20]. They propose using a particle filter that uses Markov state transition and sensor use models learned from short term training data, obtained using a tag and track approach or manual labeling. The main advantage of this approach is that the simple single-pixel sensors are cheap and easy to install, and are not perceived to be invasive or inconvenient. However, an important drawback of this approach is low accuracy: this system was reported to have 70% accuracy when tracking 3 residents over a week-long period, and in our own deployments in 3 multi-resident homes, we observed this approach to have accuracies of 65-75%. These accuracy rates may be reasonable for some applications, but confusing the identities of residents more than a third of the time could cause problems for some smart home applications such as medical monitoring. To increase identification accuracy, additional biometric sensor data, as discussed in this paper, can be included in the STAR particle filter.

Several systems including the Active Floor and the Smart Floor **instrument the floor** to locate and identify individuals [9,6]. Jenkins et al, in 2007, studied the effectiveness of using resident mass, derived using force plate signals, to identify multiple individuals in a large population [9]. Gait analysis can also be used to differentiate individuals from instrumented floors. This type of single-pixel mass-based identification approach has the advantage that it can be performed without inconveniencing the user or violating resident privacy. Existing force

plates and smart floors require careful installation to improve aesthetic appeal and achieve user acceptance; accurate force plates are also very expensive. However, more compact, cheap designs of weight sensors that have the same form factor as a floor mat can be explored for easier installation and better aesthetic appeal in the home.

Height is a weak biometric that is often used on driver's licenses or police reports, and it can not be used to definitively identify individuals from a large population. Some existing systems use invasive video cameras to identify height [5]. Nishida et al [14] propose instrumenting the entire ceiling with a dense set of ultrasonic devices to perform fine-grained location tracking with an ultrasonic radar system. However, this system is not evaluated for its ability to differentiate or identify individuals, and this approach would involve substantial deployment effort. In 2006, Jenkins et al [10] proposes placing infrared or ultrasonic distance detectors on top of doorways for identification based on height (in a poster), using an approach similar to that described in this paper. However, height sensors are not experimentally evaluated for accuracy, multiple readings are not combined as the user walks through the home to improve accuracy, and the poster does not analyze the wider ramifications of height sensing on *in-home* resident identification. To the best of our knowledge, our work is the first to analyze how height sensing can potentially be used to effectively address the multi-resident room location and identification problem in homes with high accuracy.

3 Overview: Sensing Height with Ultrasonic Sensors

To identify residents as they move throughout a home, we deploy an ultrasonic distance sensor above every exit and entry into a room. We used an off-the-shelf ultrasonic distance sensor [3] shown in figure 1a. This distance sensor sends out ultrasonic pulses at 50KHz in a diverging cone 15 to 20° off the axis of the beam. The device then measures the time taken for the echo to return, and uses it to calculate the distance to any obstacle in front of it. Only the minimal distance of any obstacle is reported. For example, when a resident stands under the device, only the distance to the top of the head is reported, while distances to the shoulder, ear etc are automatically filtered out.

Figure 1b shows the example data from the distance sensor as a subject walks repeatedly under the ultrasonic sensor mounted on top of a typical doorway eight times. The **default distance** reported is 2.1m, which is the distance when there is no obstacle in front of the ultrasonic sensor. When a subject walks under the sensor, we see minimal peaks that correspond to the ultrasonic beam making contact with the subject's head; the difference between this minimal distance and the default distance of 2.1m returns the *apparent* height of the person as she walks under the doorway. In our controlled experiments, described in the next section, we observe that this *apparent* height measured while walking, is on average less than the *erect* height measured while standing, by about 1-3cm.

Our algorithm to extract height events and height values is as follows. We first compute timestamps when the reported distance is below the default

distance with no obstacles. We then cluster these timestamps using the DBSCAN clustering algorithm [7] to compute discrete **height events**, that correspond to residents passing by or standing under the sensor. This clustering process eliminates most noise due to a single, spurious reading. Then, for each cluster of low readings, we find the minimum distance reported (i.e. the maximum height value measured). We subtract that measurement from the default height measurement with no obstacles and use the result to be the height measurement for that height event.

To *identify* residents based on measured height values, we use a Maximum Likelihood Estimate (MLE) classifier to assign each height event to one of multiple candidate residents in a home. For each height event, the MLE classifier simply computes the probability that each resident triggered it, based on the height of that resident and the error distribution of the sensor, and assigns the height event to the resident that maximizes the likelihood of the observed measurements. In the next section, we collect height data from 20 test subjects using controlled experiments in a lab to characterize the error distribution of height measurements, under diverse scenarios of passing through a doorway.

4 Experiments in a Controlled Lab Environment

4.1 Experimental Setup

We characterize the error distribution of height sensors in a controlled lab setting by placing the ultrasonic sensor on top of a doorway about 90 cm wide and having 20 users with known heights pass beneath the sensor in a controlled manner. We chose a doorway of this width because it matched the width of many of the doorways seen in our real home deployments. We selected 20 subjects of differing heights for our experiment. The distribution of heights among the 20 subjects can be inferred from any of the scatter plots in figure 2. The subjects were randomly chosen from a pool of graduate students from 20-30 years of age; 16 of our subjects were male and 4 were female.

For each subject, we first manually measured the height while standing using a tape measure. We then measured the height reported by the ultrasonic sensor when the subject stands still exactly under the sensor. The subject then walked under the sensor several times as we varied the configurations of our doorway and requested changes in the direction and speed of walking. In particular, every subject (1) Walked 20 times in a simulated narrow doorway measuring 75 cm in width, (2) Walked 21 times under the full doorway 90 cm in width (7 times perpendicular to the plane of the doorway, and 7 times on two perpendicular planes at an angle of 45° to the plane of the doorway, for a total of 21 times). We repeated the above experiments with and without shoes for each resident.

4.2 Evaluation Results

Figure 2a illustrates that, when residents are standing erect beneath the sensor, the average error across all 20 subjects is only 0.2cm, and the maximum

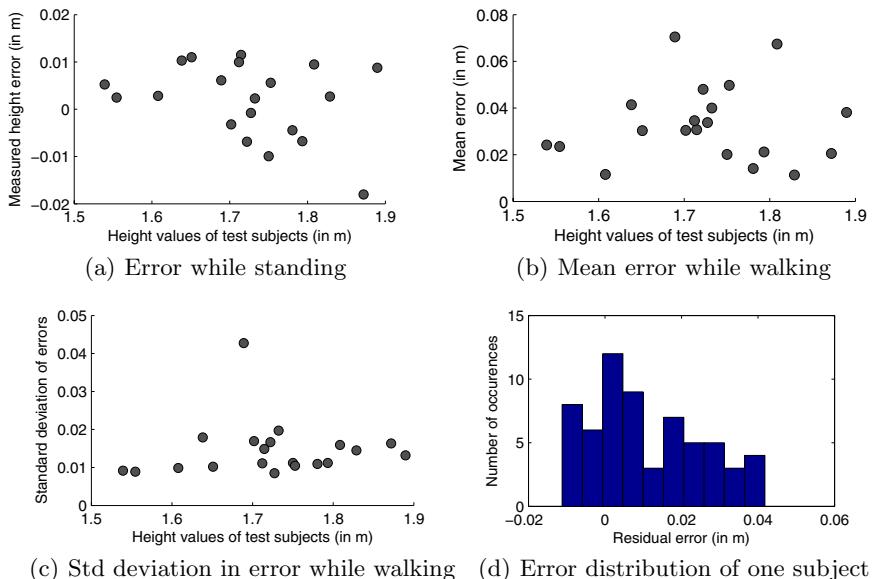


Fig. 2. Controlled laboratory experiments indicate low measured error while standing. Mean error while walking is higher due to a natural reduction in height compared to standing erect, and different walking styles. Standard error while walking is very low. The error distribution approximates a log normal distribution.

height error is 1.15 cm. Figures 2b and c show scatter plot of mean height measurement error and standard deviation in error while *walking*. *Error* here refers to the difference between the manually measured height and the height output by our height based identification algorithm for each height event. The results shown in figure 2 use the aggregated height data from all our walking experiments without shoes. We do not include results with shoes here, but observe that the mean measured height simply increased by the height of the shoe on average, and changes to the standard deviation of errors were negligible with shoes on.

From figure 2b, walking height as measured by our sensors is lower than erect, standing height by 3.31 cm on average across all subjects. This is possibly due to the natural decrease in *apparent height* as a person walks. Also, different walking styles such as bending and keeping heads down contribute to this decrease compared to erect standing height. More important is the *deviation* in residual error, the standard error, for each subject across different height events, since this will be crucial in determining identification accuracy among multiple residents in a home. We note that the mean deviation in error is only 1.45 cm. This low deviation implies that 99% identification accuracy can be obtained as long as the heights of two residents are 7cm apart. We explore this tradeoff more fully in section 7, when we describe our history based tracking algorithm using height.

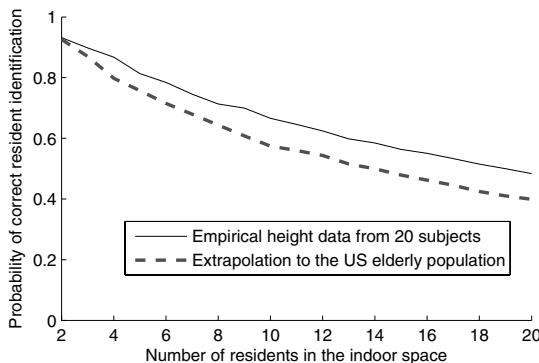


Fig. 3. Height measurements become less effective for biometric identification as the pool of individuals increases. The heights in our study were easier to differentiate than those of the general population.

In figure 2d, we show an example distribution of residual error from one test subject. The distribution shown here suggests a log normal distribution rather than a normal distribution for residual error. Thus, we ran hypotheses tests at .03 significance level for all subjects to test two different hypotheses (a) distribution of height values is normal (b) distribution of height values is log-normal. *The proportion of subjects for which the normal and log-normal hypotheses could not be rejected are 75% and 85% respectively.* The log-normal distribution, which skews naturally to the right, appears to be a better fit for modeling measured height. This is because the apparent height of a person very rarely *increases* (perhaps due to thick shoes) while walking, but more often decreases in experiments due to the ultrasonic beam making contact with the person's side (e.g.) shoulder or ear) instead of the head.

Using the empirical height data collected in the lab, we compute how well height can differentiate among a fixed set of N residents in an indoor space. In particular, we empirically calculate the accuracy with which height events are assigned to their ground truth test subjects. In figure 3, we show how this empirical identification accuracy using height decreases as we increase the number of residents under consideration in the indoor space; we randomly choose residents from our pool of 20 subjects, and evaluate how accurately individual height events generated by our subjects are labeled using a log normal MLE classifier trained from the controlled experiment data. For each N value, we repeat the random sampling 100 times. Also shown in figure 3 is how this analysis extrapolates to a national level. We model the mean and standard deviation in residual height error as a function of height using two types of curve fit models: simple linear curve fitting, and nearest neighbor interpolation. Thus, given the height of a resident, we can derive his/her mean and standard deviation, and use this in turn to derive the mean and deviation of the corresponding log normal distribution of measured height.

The 2006 health and retirement study [1] (HRS 2006) contains height measures of 4154 elderly residents living in multi-resident households. We randomly sample a fixed number N from this set of residents to be identified in an indoor space. We then *analytically* calculate the probability that any height event is assigned to the correct resident among the N residents, assuming a MLE classifier that uses the log normal distributions derived from our nearest neighbor and linear fit models. In figure 3, we show how this probability of correct event labeling degrades as we increase the number of residents in the home, randomly sampling 100 different sets of elderly residents for each N value. We only show the results with the nearest neighbor curve fit model, since only negligible differences were observed when the linear fit model was used.

As we can see in figure 3, for indoor spaces strongly resembling 2 or 3-resident elderly homes, height based identification has a mean accuracy of 87-92% using both empirical data and extrapolation to the national level. In particular, we note that 99% of the elderly multi-resident households with valid height measurements in the national study were 2-resident households. We use this insight in the next section to demonstrate the high accuracy of height based identification in 3 real multi-resident home deployments. For households with 4 residents, the identification accuracy drops to 77%; in section 7, we show that by using the history of height events at multiple doorways in a home, we can improve the identification accuracy in even 4-resident homes to 90%.



Fig. 4. Height sensors deployed above doorways in a home

5 Experiments in a Natural Home Environment

5.1 Deployment Details

Our controlled experiments characterize the sensitivity of height measurements to various conditions, including walking or standing residents, and the effect of shoes. However, these experiments do not reveal the frequency with which these conditions actually occur and affect the measurements in a natural home environment. To evaluate the accuracy of height based identification in such an environment, we deployed ultrasonic sensors in three homes for five days each. The ultrasonic sensors were deployed on doorways of rooms similar to our controlled experiments, and can be seen in figure 4. In addition to the ultrasonic sensors, we deployed the motetrack indoor localization system [13] in all homes to get ground truth locations of residents. Motetrack is a tag and track approach to localization that requires each resident to carry a mote. It uses trained RSSI signatures from beacon nodes, like the one shown in figure 5a, to localize the mobile motes in the home.



Fig. 5. Motetrack tags and beacons (left) were used to collect ground truth locations. Motion and magnetic reed switch sensors (middle/right) were used to evaluate STAR.

The main goal of the natural home deployments is to evaluate the ability of height sensors to label room visits of residents in the home, *and* compare it to a state of the art non-invasive multi-resident tracking solution that *only* relies on simple activity models derived from labeled binary sensor data [20]. In order to make this comparison, we also deployed anonymous X10 motion sensors in every room, and X10 switch sensors on daily-use objects such as the fridge, microwave, stove etc. Figures 5b and c show examples of the motion and switch sensors used in the homes. In section 6, we also compare with another well studied non-invasive resident identification solution [9] that uses resident mass to differentiate between residents in a home.

Table 1 shows some of the deployment details for the 3 homes, including number of rooms, and ground truth height values of the couple living in each home. Given the large differences in height values in the three homes, we expect our height based identification solution to perform with high accuracy.

5.2 Room Occupancy Identification in a Natural Home Environment

We evaluate the accuracy with which biometric height sensors are able to identify room visits of residents in a home. We compare the accuracy of our approach with a state of the art passive identification technique based only on ‘biometrics’ of simple activity models of residents, derived from labeled binary sensor data, as evaluated by Wilson et al in their STAR approach [20].

First, we temporally cluster X10 motion sensor firings from the same room using db-scan [7], to identify discrete room visits of residents in the home; we assume here that these temporal clusters correspond reliably to ground truth room visits of residents. Ground truth resident labels for the temporal clusters are obtained from motetrack’s location trace. Our aim is to assign resident labels to each of these clustered room visits, using either biometric height sensors, or using the location trace for each resident computed in STAR using only activity models of residents. To assign resident labels to room visits using only height sensors, we run the log normal MLE classifier on each height event that occurred during the temporal cluster. When the MLE classifier assigns a height event to a resident, that resident is added to the list of labels for that temporal cluster.

The STAR resident tracking system proposed by Wilson et al [20], uses individual Markov state transition and sensor observation models of residents to

track their activities and locations. The essence of their tracking approach is that individual residents have different movement/activity patterns in the home, and/or have unique sensor use patterns. Similar to their original implementation [20], we simply restrict our state space to include current room location of individual residents. The state transition and sensor observation probabilities are learned using counting from our ground truth training data obtained from the motetrack location trace; we performed leave-one-out cross validation over the 5 days of room occupancy data obtained from each home, i.e. for each day, we tracked residents using Markov models trained from all the other days' data. We implemented a multi-hypotheses tracking solution to track room visits of multiple residents in the home, similar to the particle filter solution implemented by Wilson et al [20].

Table 1. Details of the 3 homes used in deployments

Home	Number of rooms	Height of resident A in m	Height of resident B in m
1	7	1.88	1.77
2	4	1.68	1.55
3	5	1.75	1.63

Figure 6 compares the room labeling accuracy of our height sensor approach and the existing approach based on activity and binary sensor use models. We see that identification based on simple activity models and binary sensor use models only achieves accuracy around 65-75% in 3 homes, while height based identification achieves accuracies ranging from 98-100%. Clearly, the activity and *binary* sensor use patterns of residents in these homes are not distinguishing enough to assign room visits with high accuracy. We do not claim here that our approach is better than STAR; instead, we simply compare with an existing instantiation of the STAR framework using only 'biometric' room transition models and binary sensor use models. Certainly, as pointed out by Wilson et al [20], by using more fine-grained sensing at a higher installation cost than our approach, it is possible to better differentiate residents even using these simple models; one could even incorporate height sensor data in the STAR particle filter.

We also observe that our height sensor based approach does not require any training phase, while any approach that depends mainly on activity or binary sensor use models requires a long training phase where ground truth locations of residents need to be collected using wearable tags, to determine the probability models used in tracking; such a training phase might also require the installation of a separate infrastructure just for tracking, although recent advances in low cost tag and track solutions [15, 21] may negate the need for such a tracking infrastructure.

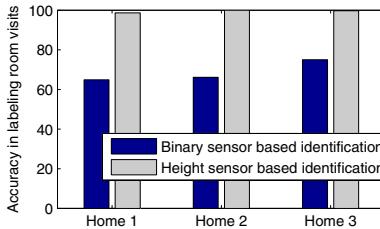


Fig. 6. Height sensors achieve higher accuracy than achieved by existing implementations of the STAR approach that use only activity models and binary sensor use models of residents for identification

6 Accuracy of Height Sensing in Homes Nationwide

We analyze the proportion of homes at a national level where our height solution can *potentially* differentiate residents with high accuracy using the 2006 health and retirement study (HRS 2006) [1], which contains height and weight measures of elderly residents living together in the same household. Of the 2107 multi-resident households with valid height and weight measures for every resident in the home, we used the 2077 households that were two-person households. We do not have currently have access to any anthropometric datasets that support our claim for a wider population demographic. We also note here that wider, longer term deployments in real homes are the best way to evaluate this technology, and our results below are best effort extrapolations from our controlled experiments in the lab.

For each home, using the height values of the residents in the home, we first derive a log normal probability model for each resident in the home using his/her height and the curve fit models described in section 4.2. We then *analytically* calculate the probability that any height event will be assigned to the right resident, assuming that each resident is equally likely to generate a height event. From now on, we refer to this probability as the **probability of correct resident identification** in a home. For each home, HRS 2006 also provides the weight measures of every resident. Jenkins et al [9] in 2007 observe that weight based identification using force plates has a Gaussian error with a mean of 0.67kg and standard deviation of 0.96. Assuming this Gaussian model and mean parameters, we calculate the probability that any gait event will be assigned to the right resident, assuming again that each resident is equally likely to generate a gait event.

Given the probability of correct resident identification for each home in the sample, we compute the proportion of homes where the probability of correct identification is above a fixed threshold; Figure 7 shows how this proportion decreases as we increase the threshold for probability of correct resident identification. Our height based identification solution is potentially applicable to 85% elderly homes in the US with at least 95% identification accuracy. Using force plates and weight based identification, up to 92% of the elderly homes can

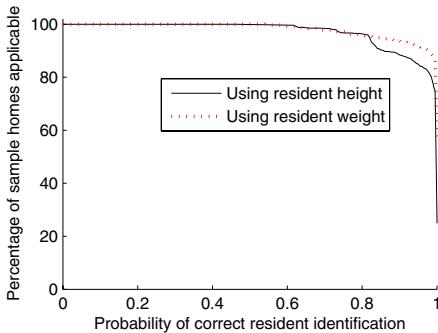


Fig. 7. Height sensors are potentially applicable with high accuracy to a large proportion of US elderly households. Weight sensors have potentially higher applicability, but require good sensor design to aesthetically install on the floor.

potentially achieve least 95% identification accuracy. Given the high cost and effort involved in retrofitting a home with force plates, height based identification is preferable, even though it is slightly less accurate; however, alternative cheaper sensing solutions for weight measurement or gait analysis can be explored for preferential use over height sensors in some homes.

7 Improving Height Measurement Accuracy with History

We have shown in the previous sections that height based biometric identification is potentially applicable to a significant proportion of elderly homes in the US with high accuracy. However, from the analysis seen in figure 7, height sensors achieve less than 95% identification accuracy in 15% of the homes. In this section, we show how information such as the room topology of a home, and the past history of height sensor events on multiple doorways in a home, can potentially be used to improve the inherent accuracy of biometric identification using the height sensor.

As an illustrative example, assume two residents A and B initially in the bedroom. Assume that after some time, resident A leaves the bedroom, and goes to the kitchen through the living room to get a snack. Even if a few *individual* height sensor events lead to incorrect results from the MLE classifier, the *sequence* of height events generated by A will have a higher likelihood of being classified as resident A; we use spatio-temporal continuity of motion through the constrained floor layout of a home to improve identification accuracy. An assumption in the in the example above, and in the analysis below, is that the error at individual height sensors is independent; this assumption may be true most of the time. However, there might occur cases where the error is more systematic, such as a person stooping over to carry a heavy object; in such cases, the utility of using the sequence of height measurements could potentially be reduced.

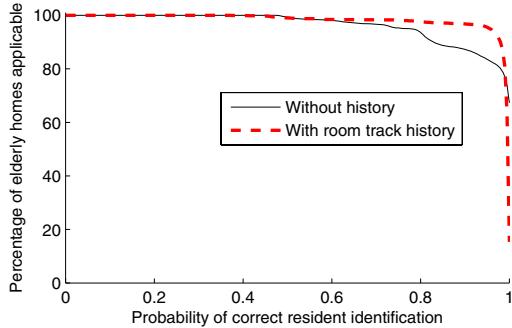


Fig. 8. Our simulation study shows that the history of height measurements collected over the track of a resident through the home potentially improves identification accuracy and applicability of height sensing in US elderly homes

We use a simulation based study that is driven by the public-use height data from HRS 2006 and our height error models derived in section 4.2, to estimate the improvement in identification accuracy that can be achieved using the history of height events in a home. We assume a 6 room home across all the elderly households for consistency. We define the same HMM for each resident's room transitions, to indicate equal transition likelihood from one room to another; we do this to ensure that differing room transition patterns of residents or specific room topologies of homes do not unfairly improve the identification accuracy possible by using the past history of height events alone. We generate 1000 height events for each home in HRS 2006, using our HMM to generate room transitions, and using the height error models from section 4.2 to generate noisy height events for each resident. We assume a height sensor at every entry/exit into a room.

Given the simulation trace, we evaluate two approaches to identify resident labels for height events in the home (1) A naive MLE classifier that only considers data from individual height events (2) A probabilistic multi-hypotheses tracker that uses past history and room topology embedded in a HMM. Figure 8 shows the applicability of height based identification across elderly homes in the US with and without history information, based on the results of our simulation experiment. When our probabilistic multi-hypotheses tracker is used, we observe that height based identification can potentially achieve at least 95% identification accuracy in 95% of elderly homes in US, as opposed to only 85% of elderly homes covered by the naive identification algorithm.

Figure 9(a) provides more insight into the scenarios where history information might be most useful. When the height difference of the residents living in the home is small, using the past history of height events greatly improves the accuracy over naive MLE classification. Using tracking history can potentially reduce the height difference required for 99% identification accuracy from 6.9cm to 3.25 cm. Figure 9(b) demonstrates another important benefit of using history, the ability to achieve higher accuracy in indoor spaces with more residents; the

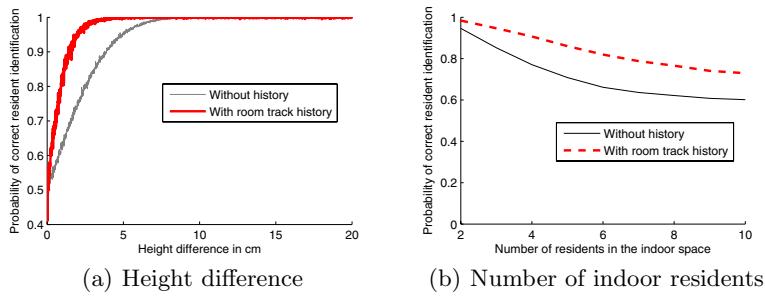


Fig. 9. A history of readings could potentially decrease the height difference required for accurate identification(a), and increase the number of residents that can be reliably differentiated(b)

heights of residents in the hypothetical multi-resident homes (100 sample homes at each point) are randomly generated using height data from the 4154 elderly residents from HRS 2006. By using the history of height events in a home, we can potentially improve the identification accuracy in 4 person homes from 77% to 90%.

8 Discussion

Height measurements can be used for accurate resident room level tracking if height sensors are placed above every entrance and exit to a room, as proposed in this paper. Of course, any biometric sensors including thumbprint or retina scanners [18], could be placed at the entrance of any room to locate residents, but violates our requirement of not requiring manual identification effort from residents. In table 2 we compare the use of height sensors for resident tracking to select existing resident tracking systems in term of four requirements: convenience, deployment time, accuracy, and cost. Deployment time is qualitatively shown for existing solutions we did not implement and is approximated from our empirical deployments in real home environments for solutions we implemented. Table 2 illustrates that, of a representative sample of four existing tracking implementations, none meet all four requirements. Tag and Track systems such as Pedestrian localization [21] and Motetrack [13] can be inconvenient to the user, STAR [20] using activity models has low accuracy and requires an inconvenient training phase, and weight sensing using force plates or smart floors [9] to track residents would require a costly installation. Of these existing implementations, only our height sensors can meet all four requirements. Of course, exploring alternative implementations of weight sensors using foot mats, or simply including height sensor data in the STAR particle filter, are possible techniques to improve existing implementations of these approaches.

In this study, we have only explored the use of height sensors above doorways to provide coarse-grained room-level accuracy. Some existing approaches such

Table 2. Evaluation of spectrum of location solutions along four variables - (Convenience, deployment time, Accuracy, Cost) assuming a five room home to deploy in

Name	Convenience	Deployment time	Accuracy	Cost
Pedestrian Localization	<i>Inconvenient</i>	Very low (5 minutes)	Very high (99-100%)	Very cheap
Motetrack	<i>Inconvenient</i>	Moderate (1 hour)	High, (95-100%)	Affordable
STAR (activity models)	<i>Inconvenient training period (weeks/months)</i>	Moderate (2 hours)	<i>Low, (65-70%)</i>	Affordable
Using Weight (force plates)	Convenient	<i>Very high</i>	High<br (>="" 95%)<="" b=""/>	<i>Very high</i>
Using Height	Convenient	Moderate (1.5 hours)	High<br (>="" 95%)<="" b=""/>	Affordable

as tag and track approaches, or invasive camera based approaches, can provide meter-level accuracy; applications that require fine-grained location accuracy would need to install height sensors *inside* rooms, such as above the stove or the sink, at a higher installation cost. An interesting research question relates to the optimal placement of height sensors inside rooms to help activity inference. A new challenge when using too many height sensors close to each other is multi-path interference affecting the ranging accuracy. Multi-path effects are also an issue in (1) rooms with wide doorways that require several adjacent height sensors to achieve sufficient coverage, and (2) adjacent doorways very close to each other; this needs to be addressed using a distributed synchronization algorithm, or careful placement of the sensors.

In our study, we use a data set which is restricted to elderly residents because it is one of the few public data sets that has both height and weight information for a large number of multi-resident homes. An interesting extension would be to explore how our solution generalizes to a larger population, including young couples, small and large families, and multi-resident student homes, by conducting large scale surveys of anthropometric measures in these homes. Since our approach is based on a resident biometric, it cannot be applied in all homes with high accuracy, unlike existing approaches such as tag and track. For applications that require higher identification accuracy than offered by height sensing alone in a given home, we propose to explore adding multiple non-invasive sensing modalities including floor mat sensor implementations for weight measurement, and color sensors above the doorway. There are also other breakdown scenarios for height sensing that we have not fully explored in this paper. The presence

of guests in the home with similar height as the existing residents will reduce identification accuracy. Also, a person who starts to use crutches or a wheelchair, might reduce identification accuracy if her new height corresponds to that of an existing resident in the home.

9 Conclusions

In this work, we demonstrate that ultrasonic range sensors placed above doorways in a home can be used to identify residents with high accuracy as they walk throughout a home, and at the same time satisfy the user requirements of smart home residents. Height is typically a weak biometric, but we make two key insights that make it effective for *in-home* monitoring. First, height is highly effective among small populations where the height differences among residents are likely to be large enough for reliable differentiation. Second, residents walk through the home in predictable, constrained patterns dictated by the floor layout, and the multiple height measurements of the resident as they walk through multiple doorways in the home can be potentially be used to improve the inherent accuracy of the height sensor. In this paper, we quantify the error with which ultrasonic height sensors measure the heights of residents as they walk under the doorway, using both controlled experiments in a lab with 20 subjects, and in 3 real homes. Using publicly available height measures of residents from multi-resident elderly homes and the height error distributions derived from our controlled and in-situ experiments, we extrapolate that a resident identification accuracy of at least 95% can potentially be achieved in 85% of elderly homes using a naive classification algorithm and in 95% of elderly homes using our probabilistic multi-hypotheses tracker.

Acknowledgments. This work was supported, in part, by the NSF grant ECCS-0901686 and the NSF grant IIS-0931972. The authors would also like to thank Timothy Hnat from the WSN group at UVa for help with our deployments, and Professor Leo Selavo from the University of Latvia for helpful initial discussions on using ranging for height sensing.

References

1. Health and retirement study (2006), <http://hrsonline.isr.umich.edu>
2. Quietcare systems - living independently, <http://www.quietcaresystems.com>
3. Vernier go motion ultrasonic sensor, <http://www.vernier.com/go/gomotion.html>
4. Wellaware systems for elderly monitoring, <http://www.wellawaresystems.com>
5. Abdelkader, B., et al.: Person identification using automatic height and stride estimation. In: International Conference on Pattern Recognition (2002)
6. Addlesee, M., Jones, A., Livesey, F., Samaria, F.: The ORL active floor. IEEE Personal Communications (1997)
7. Ester, M., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: International Conference on Knowledge Discovery and Data Mining (1996)

8. Gao, G., Whitehouse, K.: The Self-Programming Thermostat: Optimizing Setback Schedules based on Home Occupancy Patterns. In: First ACM Workshop on Embedded Sensing Systems For Energy-Efficiency In Buildings (2009)
9. Jenkins, J., Ellis, C.: Using ground reaction forces from gait analysis: body mass as a weak biometric. In: LaMarca, A., Langheinrich, M., Truong, K.N. (eds.) Pervasive 2007. LNCS, vol. 4480, pp. 251–267. Springer, Heidelberg (2007)
10. Jenkins, J., et al.: Weakly identifying system for doorway monitoring. Duke Fontiers Poster Session (May 2006)
11. Klasnja, P., Consolvo, S., Choudhury, T., Beckwith, R., Hightower, J.: Exploring privacy concerns about personal sensing. In: Proceedings of the Seventh International Conference on Pervasive Computing, Nara, Japan (May 2009)
12. Köhler, M., Patel, S., Summet, J., Stuntebeck, E., Abowd, G.: TrackSense: Infrastructure free precise indoor positioning using projected patterns. In: LaMarca, A., Langheinrich, M., Truong, K.N. (eds.) PERVASIVE 2007. LNCS, vol. 4480, pp. 334–350. Springer, Heidelberg (2007)
13. Lorincz, K., Welsh, M.: MoteTrack: a robust, decentralized approach to RF-based location tracking. Personal and Ubiquitous Computing, 489–503 (2007)
14. Nishida, Y., Murakami, S., Hori, T., Mizoguchi, H.: Minimally privacy-violative human location sensor by ultrasonic radar embedded on ceiling. In: Proceedings of IEEE Sensors (2004)
15. Patel, S., Truong, K., Abowd, G.: Powerline positioning: A practical sub-room-level indoor location system for domestic use. In: Dourish, P., Friday, A. (eds.) UbiComp 2006. LNCS, vol. 4206, pp. 441–458. Springer, Heidelberg (2006)
16. Shakhnarovich, G., et al.: Integrated face and gait recognition from multiple views. In: Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, Los Alamitos (2001)
17. Shnayder, V., et al.: Sensor networks for medical care. In: Sensys. ACM Press, New York (2005)
18. Smith, A.: Exploring the acceptability of biometrics and fingerprint technologies. International Journal of Services and Standards (2005)
19. Want, R., Hopper, A., Falcao, V., Gibbons, J.: The active badge location system. ACM Transactions on Information Systems, TOIS (1992)
20. Wilson, D., Atkeson, C.: Simultaneous tracking and activity recognition (STAR) using many anonymous, binary sensors. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) PERVASIVE 2005. LNCS, vol. 3468, pp. 62–79. Springer, Heidelberg (2005)
21. Woodman, O., Harle, R.: Pedestrian localisation for indoor environments. In: Proceedings of the 10th international conference on Ubiquitous computing, pp. 114–123. ACM, New York (2008)

Supporting Energy-Efficient Uploading Strategies for Continuous Sensing Applications on Mobile Phones

Mirco Musolesi¹, Mattia Piraccini², Kristof Fodor³, Antonio Corradi²,
and Andrew T. Campbell⁴

¹ School of Computer Science, University of St Andrews, United Kingdom

² DEIS, University of Bologna, Italy

³ Ericsson Research, Hungary

⁴ Department of Computer Science, Dartmouth College, New Hampshire, USA

Abstract. Continuous sensing applications (e.g., mobile social networking applications) are appearing on new sensor-enabled mobile phones such as the Apple iPhone, Nokia and Android phones. These applications present significant challenges to the phone's operations given the phone's limited computational and energy resources and the need for applications to share real-time continuous sensed data with back-end servers. System designers have to deal with a trade-off between data accuracy (i.e., application fidelity) and energy constraints in the design of uploading strategies between phones and back-end servers. In this paper, we present the design, implementation and evaluation of several techniques to optimize the information uploading process for continuous sensing on mobile phones. We analyze the cases of continuous and intermittent connectivity imposed by low-duty cycle design considerations or poor wireless network coverage in order to drive down energy consumption and extend the lifetime of the phone. We also show how location prediction can be integrated into this forecasting framework. We present the implementation and the experimental evaluation of these uploading techniques based on measurements from the deployment of a continuous sensing application on 20 Nokia N95 phones used by 20 people for a period of 2 weeks. Our results show that we can make significant energy savings while limiting the impact on the application fidelity, making continuous sensing a viable application for mobile phones. For example, we show that it is possible to achieve an accuracy of 80% with respect to ground-truth data while saving 60% of the traffic sent over-the-air.

1 Introduction

Over the last few years, we have witnessed the growth of personal sensing applications based on inference of human behavior and their surroundings using commercial mobile devices with on-board sensors (e.g., accelerometer, digital compass, microphone, camera). Mobile sensing applications [15][12] are being developed for new sensor-enabled mobile phones (e.g., Apple iPhone and Nokia N95) and new sensing approaches are emerging based on participatory [17] and people-centric sensing [4] paradigms, where people carrying sensor-enabled mobile phones are central to the sensing process (i.e., they are active producers and consumers of sensed data). A wide set of sensing systems are envisioned where phones are used not only to retrieve presence information about

individuals, but also to sense external environmental conditions in real-time, such as traffic, road conditions and air quality [16][7]. This new sensing area is likely to see a significant increase over the next decade with applications in both personal as well as public sensing emerging [4].

However, the development of these platforms presents a number of important design challenges particularly in terms of the availability of limited resources, such as computational capabilities and battery power. This is particularly problematic for a new class of *continuous sensing applications* found on mobile phones [4], which continuously make inferences about people and their environment and communicate sensed data in real-time with back-end server over cellular or WiFi networks. This work is based on measurements performed using phones exploiting GPRS connectivity, but the techniques described in this paper can also be applied to the case of WiFi connectivity. These resource-demanding sensing applications include social networking systems reporting user presence information such as CenceMe [15], which supports the inference of the current activity of the user carrying the device (such as sitting, standing, walking, driving, etc.) (Figure 1a). The user's sensing presence is sent from the mobile phone to social networking applications such as Facebook, MySpace and Twitter. Another example of continuous sensing application is the on-line mapping and rendering of human activities in virtual worlds such as Second Life [18] (Figure 1b) where activity performed by an individual in the physical world is mapped in real-time into actions displayed by an avatar in the virtual world.

The communication cost of continuous sensing applications is significant and can quickly lead to battery depletion. In fact, it has been shown in [15] that these continuous sensing applications using the GPRS wireless network only last a few hours. In addition, the financial cost of continuously using the wireless network may also limit the widespread deployment of these applications. Therefore, we argue that there is a need to reduce the cost related to data transmission of using these sensing applications on mobile phones: we propose a number of strategies for intelligent data uploading from mobile phones by reducing the number of transmissions. One important challenge when attempting to reduce the uploading duty cycle is that applications that require near real-time updates of the sensed data should be able to operate with missing information in a seamless way, without significantly disrupting the application fidelity. This presents a trade-off between information availability and accuracy, but in this case, the sensing system should be designed to guarantee a satisfactory user experience. In the case of the continuous sensing applications discussed above, if the information about the current state of the user is not available, a consistent state should be displayed. Since most of these applications are recreational (such as social applications), perfect accuracy is not strictly necessary.

Given these challenges, we propose to analyze the streams of sensed or inferred information on mobile phones in order to upload new data only if necessary (e.g., if the state of the user has been different from that of the back-end server for a certain period of time). Prediction mechanisms based on the past history of the user's state can be implemented on the back-end server in order to show a meaningful state if no updates have been received because of the mobile phone's low-duty cycle update strategy (i.e., updates are sent periodically and only when necessary to drive down energy costs



Fig. 1. a) CenceMe screenshots showing user activities inferred by the system. b) Avatar in Second Life: the action performed by the avatar has to be refreshed in a consistent way also in presence of disconnections.

associated with communicating with the back-end). In essence, mobile clients and the central server can be coordinated by designing predictors that receive information from the phones only if necessary.

In this paper, we present the design of the low-duty cycle uploading algorithms that consider different aspects of the accuracy/power consumption trade-offs in support of continuous sensing applications. We discuss the implementation and experimental evaluation of these mechanisms by means of measurements from the deployment of CenceMe, a sensing system based on mobile phones. We consider the case of inference of human activities, but the proposed algorithms can be directly applied to other high-level information, i.e., it is possible to exploit uploading strategies represented by means of a set of discrete states. Previous work focused on smart techniques for uploading location information [10,21]. To the best of our knowledge, this is the first work that targets the problem of devising intelligent uploading techniques of generic sequences of discrete data for sensing systems based on mobile phones. We design techniques that can operate in the following scenarios: i) connectivity is always available; ii) connectivity is intermittently available (because of duty cycle design choices in order to save energy or radio coverage); and iii) GPS information is available on the devices (assuming at least intermittent connectivity). The contributions of this paper are as follows:

- We discuss several techniques for intelligent uploading of discrete sensed information when connection is available: the key idea is to analyze and optimize the stream of states (activities) to be uploaded in order to reach an acceptable trade-off in terms of accuracy and energy consumption given the requirements of the sensing systems.
- We present an uploading strategy based on prediction mechanisms to deal with voluntary (i.e., duty cycling performed in order to save battery) and involuntary (i.e., poor cellular coverage) disconnections. More specifically, we discuss a server-side prediction algorithm to reconstruct the current user activity based on a simple, but, at the same time, effective Markov model [3] representing the probability of transitions between different states. A predictor is used in the back-end server to forecast the current state when fresh information is not present. Periodically, updates are

sent to the back-end server if necessary. The fresh information is sent if and only if the server information diverges from that currently calculated in real-time on the mobile phones. We show that it is possible to achieve an accuracy equal to about 80% with respect to the ground-truth data extracted by means of the classifiers while saving 60% of traffic sent.

- Finally, we show how location information can be used to optimize the uploading process. We observe that different user behaviors in terms of activity transitions are coupled to different geographical areas. Therefore, it is possible to associate a state transition matrix to different locations in the geographical space.

We consider the three scenarios listed above separately, but it is possible to design systems combining the proposed techniques, since they are orthogonal in many aspects. For example, a system might use the techniques based on location information when available and exploit the others, when, for example, the GPS signal is not present, because users are indoor. The proposed techniques are independent from the underlying activity recognition algorithm. It is worth noting that the aim of this work is to provide a generic framework to evaluate the trade-offs between uploading frequency of sensed data and information accuracy without considering external knowledge such as the fact that an activity is more probable in a certain area (e.g., dancing in a disco club) or that a sequence of activities is more likely than others (i.e., we do not consider semantic strategies). Our focus is on the energy consumption related to data transmission over the cellular network and not on the sensing process that can also be optimized, but this is an aspect that is also orthogonal to the techniques we present in this paper.

All these techniques have been implemented and evaluated experimentally using traces collected by distributing 20 Nokia N95 phones running the CenceMe system to university students and staff during 2 weeks over the summer. The dataset includes GPS coordinates, raw accelerometer data and inferred user activities. This represents a unique dataset containing information not only about user locations but also user activities.

2 Dataset Description

We now describe the dataset used for the proof-of-concept experiments in details. The dataset was collected during the deployment of a modified version of the CenceMe application [15] that logged all the sensed information and high-level inferred activities on the phone's on-board flash memory. The data were collected by means of 20 Nokia N95 phones carried by students and staff members from the departments of Computer Science and Biology at Dartmouth College. The dataset includes the following information for each user: accelerometer raw data, high-level activities inferred by the classifier running on the CenceMe clients, and GPS location coordinates. The dataset is available for download from the CRAWDAD website [1]. The duration of the experiment was 2 weeks. These data are used as ground-truth for our experiments, in particular, for the evaluation of prediction techniques.

The accelerometer daemon that accesses the sensor hardware (and the related classifier) has a duty cycle of 8 s (4 s sampling period and 4 s waiting time). The 4 s waiting time was introduced in the design of the system to allow for the transmission of the data

to the remote back-end server. The GPS duty cycle chosen for the experiment was 3 minutes. By doing so, it was possible to reach an acceptable compromise with respect to the accuracy/battery consumption trade-off. In fact, users had on average enough battery to run the application during the day and recharge the phone at night.

We are aware that the results presented in this work are related to the specific scenario of students and staff living in a geographical area composed of small cities, but, to the best of our knowledge, there are no other publicly available datasets with the same characteristics. However, we conjecture that this dataset can be considered as representative of activity and movement patterns of individuals in similar deployment scenarios such as campuses and communities that live in geographical areas of similar size.

3 Optimizing User State Uploading

In this section, we firstly describe the problem of state uploading and then we discuss and evaluate some algorithms for scenarios characterized by continuous and intermittent connectivity.

We model the sensing problem as follows. The inference algorithms running on the phones generate a set $\mathcal{S} = \{s_1, s_2, \dots, s_n\}$ of high-level states s_i from processing the raw sensor data. Each user/device produces a stream of data with values in the set \mathcal{S} that have to be uploaded to the back-end. In the experiments discussed in this work we consider the following set of activities $\mathcal{S} = \{\text{Sitting}, \text{Standing}, \text{Walking}, \text{Running}\}$.

We consider two cases:

- Network connectivity is always available (i.e., intermittent disconnections and transmission errors are negligible and do not affect the uploading of the information to the back-end servers), therefore *on-line* strategies *can* be used;
- Network connectivity is intermittently available (because of poor radio coverage, etc.), therefore *off-line* strategies *must* be used.

We note that the techniques used for the case of intermittent connectivity can be applied to scenarios characterized by always-on connectivity since these can be considered as limit cases of the former.

The evaluation of the techniques presented in this work are carried out considering the overhead associated to the update of the information and not the energy consumption for a specific mobile device. At the same time, we are aware that one possibility is to exploit the fact that the GPRS network interface is not powered down immediately after data transmission, so it might be convenient to send bursts of data. However, for this specific case of continuous sensing of discrete data, the generation rate of new high-level information from the on-board classifiers may not be sufficiently fast. Figure 2 shows the energy consumption profile related to the transmission of 100 bytes using a Nokia N95 (corresponding to the upload of the information related to a single user activity using XML-RPC). The figure is obtained by means of the Nokia Energy Profiler tool [20]. We observe that after the transmission of the data the interface is still powered up for less than 5 s; this interval varies for different devices and is not standardized. A 5 s interval is not enough for collecting a sufficient amount of data for the CenceMe activity classifier. In general, the transmission of information with such a degree of granularity is not required for many applications, especially recreational ones.

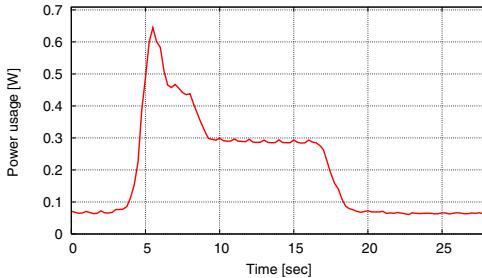


Fig. 2. Energy consumption profile related to the transmission of 100 bytes using a Nokia N95 over GPRS

3.1 Online Stream Analysis Strategies

Overview. We now present the techniques that can be used when network connectivity is always available based on the analysis of the data streams. We identify four different on-line strategies for uploading, starting from basic to more complex and optimized ones:

Always upload. The simplest solution is to upload the state of the user periodically, regardless if a change has taken place or not. This solution does not require to store any state on the mobile clients and back-end servers. It is the case without optimization. It provides 100% accuracy in the unrealistic case of no disconnections and transmission errors.

Upload in presence of changes. This strategy can be considered as an obvious optimization of the previous simple solution. The new information is uploaded every time a change takes place. This technique is the best in terms of overhead when 100% accuracy has to be guaranteed.

Upload in presence of persistent changes. According to this strategy, the new information is uploaded only when a change is not isolated, i.e., we observe a change from state A to state B with n consecutive occurrences of state B in the stream. For example, we upload the new state only after observing a sequence like $\{A, B, B, B\}$ in the case with $n = 3$. The new information is uploaded only after the n^{th} occurrence of state B . This technique involves a certain degree of information loss, since only a percentage of the actual state changes are uploaded. At the same time, this technique can be considered as a way of filtering out outliers from the data stream.

Voting based uploading strategy. This method is based on the evaluation of the frequency of activities in the data stream considering non overlapping time windows. The state with the highest frequency in the window is selected for uploading. The update is sent only if the most frequent state in the current window is different from the most frequent state in the previous window. Let us consider the following example. Let us assume that we have the sequence $\{A, A, B, A, B, A, A, B, B\}$ with window size equal to 3 and a threshold equal to 2. The first state to be uploaded is A . Then, no upload takes place in the following window, since the state with the highest frequency is still A . Finally, since the state B has the highest frequency inside the window, the update is sent to the back-end.

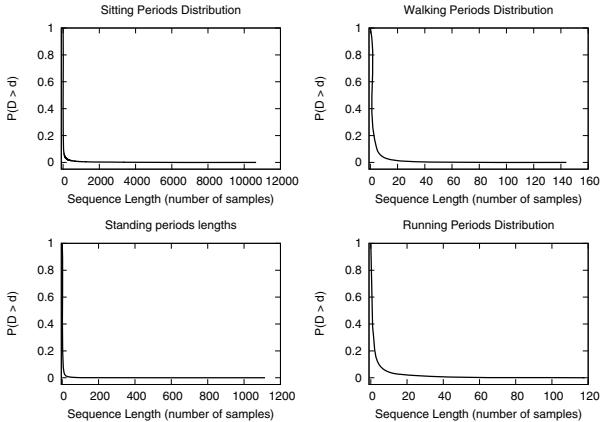


Fig. 3. Complementary cumulative distribution function indicating the probability of having a sequence of the same activity longer than Sample Length

Compression algorithms [14] can improve system performance but we do not consider them in this work, since these techniques can be easily added on top of the uploading mechanisms discussed in this paper.

Evaluation. We compare the accuracy and transmission overhead of all the techniques with respect to the *upload in presence of changes* strategy. We define accuracy as the ratio of correctly predicted values on the server against the ground-truth state inferred on the phone. The analysis performed in this section can be considered as a general methodology to be used in order to set the parameters of the algorithms in different practical cases. The results obtained in this analysis are specific to the dataset collected through the deployment of CenceMe, but the evaluation process itself can be applied to other deployment scenarios.

We first present a statistical description of the types of activities in the stream of data. In Figure 3 we show the probability of having n consecutive activities of the same type. As the plot shows, the presence of very long sequences of consecutive activities of the same kind is unlikely. These graphs give other interesting information about the length of the sequences of consecutive values of the same activity: for all types of activities in our stream, we observe the length of the sequence of the same activity is one for more than 50% of the samples. The choice of the window length parameter is a key aspect of the uploading strategy. Figure 3 shows that there is a very low probability of having long sequences of the same activity within the data stream. The probability of having sequences of the same activity shorter than or equal to 30 samples is 99% (for all the activities except sitting for which the probability is 90%). Thus, it is useless to apply filters with a window length longer than 30 samples because, in that case, the majority of the changes in the stream would be missed.

Figures 4a and 4b show the accuracy of the two methods and the ratio of the traffic sent with respect to the overhead associated with the *upload in presence of changes* strategy. In this case the number of the required consecutive changes for an uploading

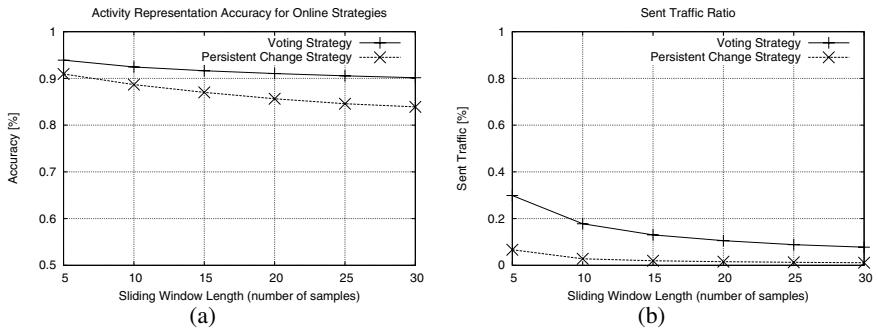


Fig. 4. a) Accuracy of the *voting based* and *persistent change* uploading strategies with respect to the overhead associated to the *upload in presence of changes* strategy. b) Traffic ratio of the *voting based* and *persistent change* uploading strategies with respect to the overhead associated to the *upload in presence of changes* strategy.

is equal to the window size. As expected, the *voting based* uploading strategy has a better accuracy than the *persistent change* strategy but it is characterized by a higher overhead. Both of them can achieve a 90% accuracy saving 80% of data traffic. We observe that the gap between the accuracy values related to the two strategies increases as the length of the window increases. The *persistence change* strategy has a smaller number of updates with respect to the voting based strategy, since the latter at each step always uploads the state calculated using the voting mechanism. The *persistence change* strategy uploads a new state only if a new state has been observed for a certain number of previous steps.

3.2 Off-line Strategies: Markov Chain Based Prediction

Overview. The strategies outlined above are based on the assumption of continuous availability of network connectivity. When the uploading strategies described in the previous section are used, the application on the phone side is responsible for choosing which state update has to be sent and when. The back-end server is not involved in the process. When the mobile device is disconnected from the Internet, the back-end can just make the last known state available or publish an *unknown state* message.

An alternative strategy is to try to forecast the next state during a disconnection. A possible cause of intermittent connectivity is insufficient radio coverage. In some cases frequent updates have to be avoided given energy constraints of the devices. This strategy can be combined with one of the mechanisms described in the previous section in presence of intermittent connectivity. In other words, online strategies can be used when connectivity is present and offline strategies when the device cannot (should not) connect to the network.

By definition, the predicted state is characterized by a certain degree of uncertainty, but this can be acceptable for some classes of systems such as recreational sensing applications like CenceMe. We are aware that the applicability of these techniques are

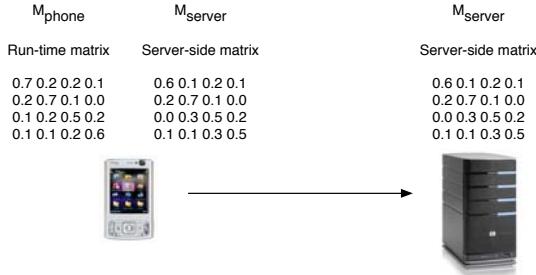


Fig. 5. Markov chain based prediction: in the mobile phones, two matrices are stored, one that is updated as new states are generated by the classifiers and a copy of the matrix that is periodically sent to the back-end server when the two diverge

not universal: examples of non viability include monitoring technologies for healthcare and assisted living [6], for which it may be necessary to guarantee perfect accuracy.

The key idea is to use a transition matrix to model the sequence of the state changes (such as sequences of actions of user avatars) on the server also during a disconnection from the mobile client. In order to do so, we exploit a simple Markov chain model to describe the phenomenon under observation, i.e., the transitions between the states that can be sensed by the system. The matrix stores the probability of transition between the different states. These probabilities are estimated by measuring the frequency of transitions. The calculation of this matrix takes place on the phones. Instead of uploading a single state as before, the phone uploads the state transition matrix that is used by the back-end to predict and publish the next state during a disconnection. Two matrices are stored on the mobile phones, one that is updated as new states are generated by the classifiers and a copy of the matrix that was sent to the back-end server. When the system is bootstrapped, the first matrix is uploaded directly to the server since no comparison is possible. A new matrix is sent to the back-end if and only if the matrix currently calculated on the phones diverges from what is currently used by the server. The comparison is based on a difference threshold. The mechanism is shown in Figure 5. We note that for applications such as visualization of human activities in virtual worlds (e.g., Second Life), this information can be used to drive the sequences of actions of the avatars also during periods of disconnection in order to provide a better user experience.

More formally, we model the system as a stochastic process $X(t)$ (with $t = 0, 1, 2, \dots$ instants of time) that takes a finite number of possible values defined by the set of states \mathcal{S} . For a Markov chain, the conditional distribution of any future state $X(t+1)$ given the past states $X(0), X(1), X(2), \dots, X(t-1)$ and the current state $X(t)$ is independent from the values of the past states and depends only on the present state [3]. We define a matrix of transitions \mathbf{M} where each element of the matrix $m_{i,j}$ represents the probability of transitions between a state i and a state j . Each matrix is built locally on the phone considering the entire set of samples collected in a certain interval $T_{\text{calculation}}$. We argue that this matrix \mathbf{M} is in general a function of the geographical location and the time of the day, more formally $\mathbf{M} = \mathbf{M}(x, y, t)$ where (x, y) indicates the geographical position and t is the instant of time or a time interval (such as mornings, or a

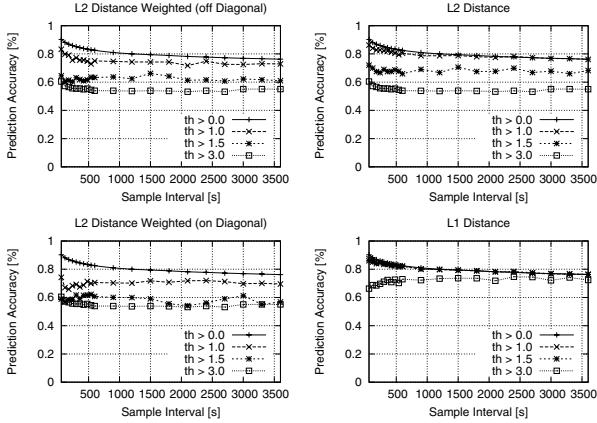


Fig. 6. Accuracy of the Markov chain model with different metrics, thresholds and sample times

particular day of the week, such as Mondays). In the next section we will discuss how different matrices can be associated to various locations in the geographical area where users move, if GPS receivers are available.

Periodically, with an interval equal to $T_{\text{calculation}} \text{ s}$, a decision step takes place: a new matrix is uploaded if the matrix on the server diverges from that currently stored on the phones. Therefore, the key problem is to measure the difference between the matrix M_{server} currently available on the back-end server and that currently estimated on the phone that we indicate with M_{phone} . In order to evaluate this error we calculate the distance between the two matrices M_{server} and M_{phone} . More specifically, a new matrix is uploaded to the back-end if and only if

$$L_x(M_{\text{phone}}, M_{\text{server}}) \geq th \quad (1)$$

where th is a pre-defined threshold and L_x is a chosen distance function. In fact, by considering a matrix as a vector, we can calculate the distance between two subsequent vectors using standard vector distances [9]. A basic choice is to use the Euclidean distance between two matrices defined as follows:

$$L_2(M_{\text{phone}}, M_{\text{server}}) = \sqrt{\sum_{i,j \in \mathcal{S}} (m_{\text{phone}_{i,j}} - m_{\text{server}_{i,j}})^2} \quad (2)$$

We also consider other two distances, the so-called Manhattan distance L_1 and the weighted distance L_{2w} . In our case, the L_1 distance is defined as follows:

$$L_1(M_{\text{phone}}, M_{\text{server}}) = \sum_{i,j \in \mathcal{S}} |m_{\text{phone}_{i,j}} - m_{\text{server}_{i,j}}| \quad (3)$$

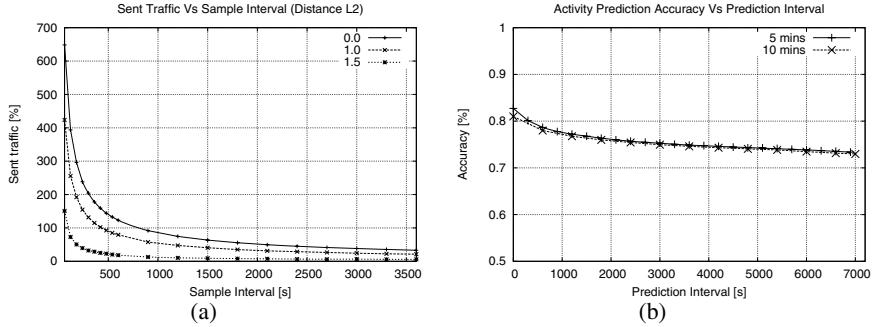


Fig. 7. a) Traffic sent versus sample interval for different thresholds in the Markov model. b) Accuracy of the Markov upload model in case of disconnection of the mobile device from the back-end.

The L_1 metric provides an approximation of the Euclidean distance but it is less expensive to compute. We use the following formula for the calculation of L_{2w} distance:

$$L_{2w}(M_{phone}, M_{server}) = \left(\sum_{i,j \in \mathcal{S}, i=j}^{|S|} (w_d(m_{phone_{i,j}} - m_{server_{i,j}})^2) + \sum_{i,j \in \mathcal{S}, i \neq j}^{|S|} (w_{nd}(m_{phone_{i,j}} - m_{server_{i,j}})^2) \right)^{\frac{1}{2}} \quad (4)$$

Using this distance, we can assign higher importance to one of the two classes of transitions described by the matrix M : self-transitions from one state to itself (self-transitions) and those from a state to a different one. The probability of self-transitions are represented by the elements of the diagonal.

Evaluation. In this section, we present the results of the evaluation of the Markov chain strategy varying both the values of the distance thresholds and sample intervals. As for all strategies, we are interested in studying its accuracy and overhead.

Figure 6 shows the accuracy of the Markov model depending on the values of the sample interval and threshold. The plots refer to the L_1 , L_2 and L_{2w} distances. All the metrics show a decrease in accuracy depending on the value of the sample interval $T_{calculation}$. As expected, the accuracy strongly depends on the value of the threshold used for the uploading decision. More specifically, we observe that when the value of the threshold is lower than 1 the accuracy of the prediction does not change. Instead, with a threshold higher than 1, the accuracy decreases because less transition matrices are sent to the back-end. At the same time, we observe that by uploading the matrix more frequently (i.e., by using a lower threshold), it is possible to achieve a better accuracy. Small sample intervals do not provide a statistically valid number of samples and, therefore, the quality of the prediction is rather poor in these cases. We plot results with thresholds up to 3, since we note that the measured accuracy does not change with a threshold greater than 2.5, a value that represents the maximum distance measured between two consecutive matrices for all users in the experiment.

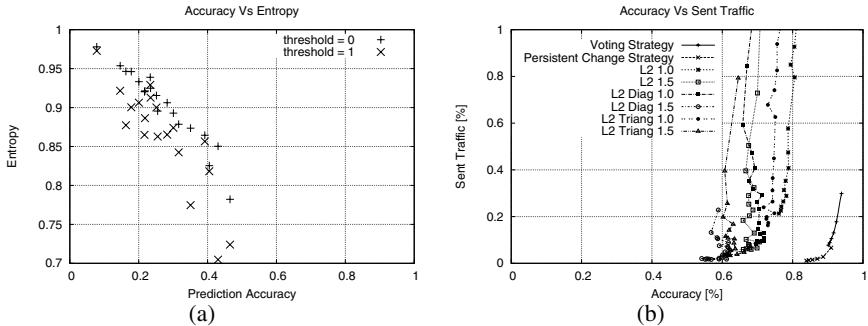


Fig. 8. a) Correlation between prediction accuracy and entropy value for each user using different thresholds using the Markov model. b) Cross-evaluation of off-line and on-line models: correlation between accuracy and sent traffic.

In general, we observe that the L_1 distance method provides the best solution in this case since it is less influenced by the choice of the parameters. As for the other strategies, we measure the difference of overhead (in percentage) with respect to the upload in presence of changes strategy. Figure 7a shows the percentage of traffic sent by this method with respect to the basic one. We would like to underline that for this method, every time the application transmits some data, instead of sending a single state, it sends $card(\mathcal{S}) \times card(\mathcal{S})$ probability values. For low values of the sample interval the amount of traffic is up to seven times higher than the basic method. This value decreases rapidly for higher sample intervals. Another parameter affecting the traffic overhead is the distance threshold. As expected, for higher values of the threshold, less matrices are sent and the amount of transmitted data decreases. As the plot in Figure 7a shows, the Markov model presents the same traffic load of the baseline model when a matrix is sent every 800, 400 and 50 s respectively for thresholds equal to 0.0, 1.0 and 1.5.

For this strategy, it is fundamental to measure the accuracy of the method in case of disconnection of mobile devices from the back-end. Figure 7b shows the decreasing accuracy of the Markov predictor as time goes on. We analyze two cases in which a matrix is uploaded respectively every 5 or 10 minutes (the two lines in Figure 7b). We build a transition matrix considering 5 and 10 minute calculation intervals, then we assume that a disconnection takes place: no more matrices can be uploaded and we keep predicting the next activity samples with the same matrix. In this case we observe about a 10% accuracy reduction.

The results presented above are derived by considering aggregated data for all users. We now present a possible method to tie the accuracy of the prediction of the activities of a *single* user to a quantitative measure. We observe that also intuitively user predictability is strongly dependent on the degree of user behavior variability: the higher the variability of activity transitions, the less predictable the user is.

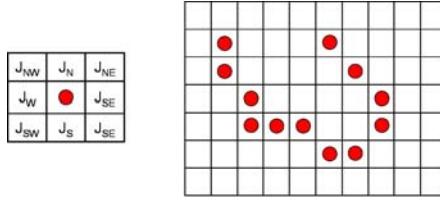


Fig. 9. Transition probability matrix used in the local model and example of a user path

A standard measure of state variability in user data streams is the entropy of the sequence; we use the standard definition provided by Shannon [22]. In Figure 8a we show the relation between entropy and the prediction accuracy described in the previous sections. Each point in the plot corresponds to a user involved in the experiment; we report the values related to the cases with thresholds equal to 0 and 1. For the first set of points, it is possible to observe a sort of linear distribution, while the other set forms a cloud-like distribution. This means that the correlation between accuracy and entropy is strictly dependent on the choice of the parameters used to tune the offline strategy. The accuracy decreases as the threshold increases and this is directly reflected on the negative correlation between entropy and accuracy.

The state entropy can provide a measurement of the predictability of a certain user and this information might be displayed together with the forecasted state when the Markov based model is used. These results also provide a statistical characterization of the dataset and can be exploited to interpret and compare performance results in this paper with user patterns in deployments related to different social scenarios.

3.3 Comparison between On-line and Off-line Strategies

Finally, we compare the results of the on-line and off-line strategies described above. This comparison is performed only for completeness, since the two strategies are targeting two different scenarios, characterized by different types of connectivity (continuous and intermittent) and/or energy requirements (i.e., battery power constraints are deemed less or more important than accuracy).

Figure 8b shows the correlation between accuracy and overhead using the two different classes of strategies, by increasing window sizes for the online strategies and different sample intervals for the Markov model based ones. The on-line strategies show the best performance: in terms of accuracy, the *voting based* strategy provides the best results (93% accuracy); the *persistent change* uploading strategy instead has the best performances in terms of overhead, i.e., it saves up to 99.9% of traffic.

As expected, the Markov model based approach cannot provide better accuracy than online strategies. However, we would like to point out that the Markov based strategy can offer an interesting trade-off between accuracy and overhead also when connectivity is available. For example, using a L_2 metric with a threshold equal to 1, it is possible to achieve an accuracy equal to almost 80% by saving 60% of traffic sent.

4 Location-Based State Uploading

We now show how location information can also be used to optimize the uploading process in case of devices equipped with GPS receivers. The key idea is to associate a state transition matrix to each location. This can be considered as a sort of augmentation of the Markov chain based mechanism presented in the previous section. It can be used when connectivity is potentially intermittent or continuous uploads are not possible given battery constraints. More specifically, we exploit a two-level Markov model: we firstly use a transition matrix associated to each location of the space to predict future movements. Then we use the matrix associated to the forecasted next location for the prediction of the future activity. The matrices are built by collecting data for a certain interval of time (that can be considered as a sort of training period of the model). Then these matrices are used for forecasting on the server if no data are transmitted by the mobile clients.

Markov models have already been successfully used as a basis of user location prediction techniques [2][23][11]. A key problem is the definition of the locations in the geographical space. In order to apply a Markov model for location prediction we need to transform the continuous domain of a geographical region into a discrete set of areas. Then, a way of estimating the probability of transitions between these areas has to be devised. We use a grid based model for subdividing the geographical areas in discrete locations, obtaining a grid of squared tiles (or cells). More formally, we divide the space in $m \times n$ squared tiles $T_{p,q}$ with side size $gridsize$. We then consider the probability of transitions between tiles by assuming two movement models with different constraints in terms of possible transitions, i.e. a *local* one and a *global* one.

Local Movement Model. We start considering a simple movement model that takes into consideration only transitions to locations that are close to the current one. According to this model, a user in a tile $T_{p,q}$ can only jump to a tile which is adjacent to that she is currently located (considering an analogy to the game of chess, the user can move as a king on a chessboard). For every tile of the matrix we consider a vector $J = \{J_N, J_{NE}, J_{NW}, J_S, J_{SE}, J_{SW}, J_E, J_S, J_{CT}\}$ containing the probabilities of jumping from the current tile to one of the tiles in the neighborhood (the north-east tile, the north tile, and so on) as illustrated in Figure 9. The matrix also includes the probability of staying in the current tile (J_{CT}). We do not simply take into account jumps to tiles that are further on the grid.

The local movement model is a simple way of representing the movements of the users, however, it is not sufficiently accurate because it does not take into account possible jumps between two non adjacent tiles. This type of jumps can be caused by various physical factors, such as the speed of users with respect to a low sampling rate (e.g., a movement in a car), the absence of GPS signal and/or the unavailability of the GPRS coverage. In these cases, some transitions might not be recorded.

The two key parameters affecting this model are GPS sensor duty cycle and size of tiles. As the duty cycle of the GPS sensor increases, the probability of a transition between two adjacent tiles decreases. A value of $gridsize$ smaller or equal to the maximum distance that a user can cover increases the probability of jumps between two non adjacent tiles. One of the advantages of the local model is related to the memory that is needed to implement it. In fact, for each tile of the area, we need to store only a

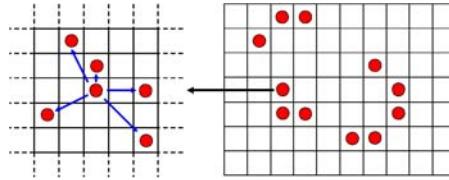


Fig. 10. The global model and a possible path described by a user. For each cell of the area $N \times M$ (on the right), a $N \times M$ matrix of transition probabilities is recorded.

3×3 matrix (or a 9-cell vector containing the probabilities of jumping to the adjacent tiles) for a total space complexity equal to $M \times N \times 3 \times 3$ where M and N are the dimensions of the geographical area of interest. If we consider the nine cells containing the transition probabilities as constant values, the space complexity of the model is approximately $O(MN)$.

Global Movement Model. We also consider a movement model that allows jumps between non adjacent tiles (global movements). In other words, the global movement model is able to register all user movements over the grid: both jumps toward adjacent tiles and tiles that are far away are allowed and recorded, as illustrated in Figure 10. However, this model requires an array of size $M \times N$ representing a matrix containing all the transitions toward all the other possible geographical locations for each tile. The drawback of using this model is the increased space complexity. In fact, for each tile of the geographical area it is necessary to store a $M \times N$ probability matrix where every cell $T_{i,j}$ contains the probability of jumping from the current tile to any other tile of the geographical area. The space complexity of the model becomes $O((NM)^2)$.

Evaluation. We use the first week as training period (i.e., the matrix is sent once after the first week) and we measure the accuracy of the prediction considering the dataset corresponding to the second week. We train the model on the entire week in order to have sufficient statistics. In a real system, an uploading mechanism based on distances as that presented in the previous section can be used for deciding if a new matrix has to be sent or not to the back-end server. The evaluation of these mechanisms is not possible for us since we have only a two-week dataset.

The main challenge in using the grid model is the definition and the placement of the grid structure. A key parameter of the model is the size of the tiles composing the superimposed grid. In fact, by using a grid, a logical place (e.g., the library, the gym, etc.) might be split into two or more tiles and two unrelated places might be unified in the same tile. Even if we keep the same grid size, by changing the offset of the grid we have very different subdivisions of the area. Accuracy variations are around 5 to 10% with different grid positioning. In Figure 11 we show the accuracy of the prediction as a function of the grid size. We plot the results for the local and the global models with different strategies for the choice of the first activity that is used as first state of the prediction model when the transition between two tiles of the grid takes place (i.e., the starting state of the activity transition matrix associated to each tile). More specifically, we use the most popular activity in that tile and a randomly selected activity using a frequency distribution of the activities in that location. These results are obtained by

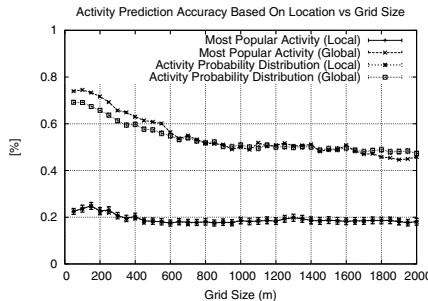


Fig. 11. Activity accuracy prediction as a function of different grid sizes

averaging 10 runs with different offsets from the origin of the grid. The offsets are equal to $\frac{k}{10} \text{gridsize}$ with $k = 1 \dots 10$. We note that this choice does not affect the performance of the algorithm as expected, given the properties of Markov chains about independence from the initial state as time passes [3]. We note that the accuracy of the prediction decreases for all the methods as the grid size increases. In fact, with a larger granularity of the tile, locations characterized by different activities are joined together (for example the library and the street leading to it). The local prediction model is not able to capture the movements of the users. These results are also affected by higher errors with small grid size for the location prediction due to a higher number of tiles with insufficient statistics. The local model is particularly affected by this problem, since many transitions to neighboring cells are not recorded.

5 Related Work

This paper proposes a set of intelligent uploading techniques for (near) real-time continuous sensing systems based on mobile phones. The existing related work focuses mainly on movement prediction and the use of information about the current location to infer human activities (such as cooking is very probable in a kitchen). The problem of optimizing the uploading process has not been explored yet also because mobile sensing systems based on smart phones, like MyExperience [8], CenceMe [15], Nericell [16] and BeTelGeuse [12], are very recent. Recently, Wang et al. have proposed a framework called EEMSS [25] for optimizing the duty cycles of the sensors for mobile sensing applications. The aim is orthogonal to ours and the solutions can coexist, since EEMSS is focused on the sensing aspect whereas the goal of this work is to optimize the uploading process.

With respect to the problem of forecasting user movements, in [2], the authors present a model of user location prediction from GPS data. A simple first-order Markov model to predict the transitions between significant places is used. Also in this work temporal aspects are not taken into consideration. Moreover, the model is not able to forecast transitions to geographical areas that are not considered significant. In [13] the significant places are extracted by means of a discriminative relational Markov network; then, a generative dynamic Bayesian network is used to learn transportation routines.

Another system for the prediction of future network connectivity based on a second-order Markov model is BreadCrumbs [19]. This system is able to predict only the next user location and not the time of the transitions and the interval of time during which users reside in that specific location. An extension of [24] about the study of handoff mechanisms in a campus environment using Markov models is presented in [23].

GPS information has also been used in vehicular systems such as in Predestination [11]. The authors of this work use aggregated location data together with additional information about the geography of certain areas in order to make accurate prediction of movements of vehicles; we have presented instead a technique that relies only on local predictions of single users also in presence of intermittent connectivity. Techniques for approximating the positions of moving objects also considering energy requirements have also been studied in [10] and [5]. Other related work has been done in the area of databases in particular about the so-called *approximation replication* techniques [21]. The key difference with respect to this body of work is related to the fact that our goal is also to provide an estimation of the current state using the past history.

6 Concluding Remarks

In this paper we have presented a series of techniques for optimizing the uploading process of discrete data for continuous sensing applications on mobile phones. We have considered two cases, namely a scenario where connectivity is always available and one where it is intermittently present or the number of transmissions has to be limited given the system design requirements in order to extend the battery lifetime of the devices. Finally, we have shown how location information can be exploited to optimize the uploading process. We have demonstrated that our techniques can be used to improve the performance of continuous mobile sensing applications by analyzing the trade-off between transmission overhead and accuracy.

Acknowledgments. We would like to thank Petteri Nurmi, our shepherd, for his comments. This work is supported in part by Intel Corp., Nokia, NSF NCS-0631289, and the Institute for Security Technology Studies (ISTS) at Dartmouth College. ISTS support is provided by the U.S. Department of Homeland Security under award 2006-CS-001-000001, and by award 60NANB6D6130 from the U.S. Department of Commerce. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of any funding body.

References

1. CRAWDAD Project, <http://www.crawdad.org>
2. Ashbrook, D., Starner, T.: Using GPS to Learn Significant Locations and Predict Movement Across Multiple Users. *Personal and Ubiquitous Computing* 7(5), 275–286 (2003)
3. Brémaud, P.: *Markov Chains, Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, Heidelberg (1998)
4. Campbell, A.T., Eisenman, S.B., Lane, N.D., Miluzzo, E., Peterson, R., Lu, H., Zheng, X., Musolesi, M., Fodor, K., Ahn, G.-S.: The Rise of People-Centric Sensing. *IEEE Internet Computing Special Issue on Mesh Networks* (June/July 2008)

5. Civilis, A., Jensen, C.S., Pakalnis, S.: Techniques for efficient road-network-based tracking of moving objects. *IEEE Transactions on Knowledge and Data Engineering* 17(5), 698–712 (2005)
6. Consolvo, S., Everitt, K., Smith, I., Landay, J.A.: Design requirements for technologies that encourage physical activity. In: *Proceedings of CHI 2006*, pp. 457–466. ACM Press, New York (2006)
7. Eriksson, J., Girod, L., Hull, B., Newton, R., Madden, S., Balakrishnan, H.: The Pothole Patrol: using a Mobile Sensor Network for Road Surface Monitoring. In: *Proceedings of MobiSys 2008*, pp. 29–39. ACM, New York (2008)
8. Froehlich, J., Chen, M.Y., Consolvo, S., Harrison, B., Landay, J.A.: MyExperience: a System for in Situ Tracing and Capturing of User Feedback on Mobile Phones. In: *Proceedings of MobiSys 2007*, pp. 57–70. ACM, New York (2007)
9. Horn, R.A., Johnson, C.R.: *Matrix Analysis*. Cambridge University Press, Cambridge (1990)
10. Kjærgaard, M.B., Langdal, J., Godsk, T., Toftkær, T.: Entracked: energy-efficient robust position tracking for mobile devices. In: *Proceedings of MobiSys 2009*, pp. 221–234. ACM, New York (2009)
11. Krumm, J., Horvitz, E.: Predestination: Inferring Destinations from Partial Trajectories. In: Dourish, P., Friday, A. (eds.) *UbiComp 2006*. LNCS, vol. 4206, pp. 243–260. Springer, Heidelberg (2006)
12. Kukkonen, J., Lagerspetz, E., Nurmi, P., Andersson, M.: Betelgeuse: A platform for gathering and processing situational data. *IEEE Pervasive Computing* 8(2), 49–56 (2009)
13. Liao, L., Patterson, D.J., Fox, D., Kautz, H.: Building Personal Maps from GPS Data. In: *Proceedings of IJCAI Workshop on Modeling Others from Observation* (2005)
14. MacKay, D.J.C.: *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge (2003)
15. Miluzzo, E., Lane, N.D., Fodor, K., Peterson, R.A., Lu, H., Musolesi, M., Eisenman, S.B., Zheng, X., Campbell, A.T.: Sensing Meets Mobile Social Networks: the Design, Implementation and Evaluation of the CenceMe Application. In: *Proceedings of SenSys 2008*, November 2008, pp. 337–350 (2008)
16. Mohan, P., Padmanabhan, V.N., Ramjee, R.: Nericell: Rich Monitoring of Road and Traffic Conditions using Mobile Smartphones. In: *Proceedings of SenSys 2008*, pp. 323–336. ACM, New York (2008)
17. Mun, M., Reddy, S., Shilton, K., Yau, N., Burke, J., Estrin, D., Hansen, M.H., Howard, E., West, R., Boda, P.: PEIR, the Personal Environmental Impact Report, as a Platform for Participatory Sensing Systems Research. In: *Proceedings of MobiSys 2009*, pp. 55–68 (2009)
18. Musolesi, M., Miluzzo, E., Lane, N.D., Eisenman, S.B., Campbell, A.T.: The Second Life of a Sensor: Integrating Real-world Experience in Virtual Worlds using Mobile Phones. In: *Proceedings of HotEmNets 2008*, Charlottesville, Virginia, USA. ACM Press, New York (2008)
19. Nicholson, A.J., Noble, B.D.: BreadCrumbs: Forecasting Mobile Connectivity. In: *Proceedings of MobiCom 2008*, pp. 46–57. ACM, New York (2008)
20. Nokia. Nokia Energy Profiler 1.1, <http://www.forum.nokia.com>
21. Olston, C., Widom, J.: Efficient monitoring and querying of distributed, dynamic data via approximate replication. *IEEE Data Engineering Bulletin* 28(1), 11–18 (2005)
22. Shannon, C.E.: A Mathematical Theory of Communications. *Bell System Technical Journal* 27(7), 379–423 (1948)
23. Song, L., Deshpande, U., Kozat, U.C., Kotz, D., Jain, R.: Predictability of WLAN Mobility and its Effects on Bandwidth Provisioning. In: *Proceedings of INFOCOM 2006* (April 2006)
24. Song, L., Kotz, D.: Evaluating Location Predictors with Extensive Wi-Fi Mobility Data. In: *Proceedings of INFOCOM 2004*, pp. 1414–1424 (2004)
25. Wang, Y., Lin, J., Annavaram, M., Jacobson, Q.A., Hong, J., Krishnamachari, B., Sadeh, N.: A Framework of Energy Efficient Mobile Sensing for Automatic User State Recognition. In: *Proceedings of MobiSys 2009*, pp. 179–192. ACM, New York (2009)

Efficient Resource-Aware Hybrid Configuration of Distributed Pervasive Applications

Stephan Schuhmann, Klaus Herrmann, and Kurt Rothermel

University of Stuttgart, Institute of Parallel and Distributed Systems (IPVS),

Universitätsstr. 38, 70569 Stuttgart, Germany

`{schuhmann,herrmann,rothermel}@ipvs.uni-stuttgart.de`

Abstract. As the size and complexity of Pervasive Computing environments increases, configuration and adaptation of distributed applications gains importance. These tasks require automated system support, since users must not be distracted by the (re-)composition of applications. In homogeneous ad hoc scenarios, relying on decentralized configuration schemes is obviously mandatory, while centralized approaches may help to reduce latencies in weakly heterogeneous infrastructure-based environments. However, in case of strongly heterogeneous pervasive environments including several resource-rich and resource-weak devices, both approaches may lead to suboptimal results concerning configuration latencies: While the resource-weak devices represent bottlenecks for decentralized configuration, the centralized approach faces the problem of not utilizing parallelism. Instead, a hybrid approach that involves only the subset of resource-rich devices is capable of rendering configuration and adaptation processes more efficiently. In this paper, we present such a resource-aware hybrid scheme that effectively reduces the time required for configuration processes. This is accomplished by a balanced-load clustering scheme that exploits the computational power of resource-rich devices, while avoiding bottlenecks in (re-)configurations. We present real-world evaluations which confirm that our approach reduces configuration latencies in heterogeneous environments by more than 30% compared to totally centralized and totally decentralized approaches. This is an important step towards seamless application configuration.

1 Introduction

The Pervasive Computing research area focuses on the development of abstractions and concepts for seamless integration of information processing into everyday activities and objects. In such environments, resources are normally scattered among the devices and any single device is not capable of executing an entire application. Thus, distributed applications need to be configured prior to their execution to ensure all required functionality is available. Configuring an application means finding a set of components which can be instantiated at the same time. Thus, application configuration is also known as *composition* of the required resources and services. Furthermore, this composition has to fulfill the structural constraints given by the required application functionalities,



Fig. 1. Distributed presentation application

while considering the limited resources in the pervasive environment. Moreover, automation is needed to make this complex process transparent for the user. As an example, consider a conference environment depicted in Figure 1 where a speaker wants to give a presentation. For this purpose, the available input resources (e.g., keyboards, microphones, touch screens) and output resources (e.g., video projectors, loudspeakers) have to be leveraged by the distributed presentation application, as demonstrated in [14]. Further typical applications provide the easy sharing of files and resources like printers or webcams with other users (Casca, [9]) or the flexible and generic control of devices and services in home media networks (OSCAR, [20]). The actual composition of the application (called *configuration*) has to be calculated by configuration algorithms on the available devices. Furthermore, automatic runtime *adaptation* (or *re-configuration*) is necessary due to the dynamism in pervasive environments. Adaptation denotes the task of finding alternative components for those parts of the application that have become invalid, e.g. due to device failures. As distractions are highly undesirable during application execution, our main goal is to perform (re-)configuration processes as fast as possible.

Two fundamentally diverse approaches for configuration and adaptation of distributed applications exist, namely *decentralized* and *centralized* configuration. Decentralized approaches focus on mobile ad hoc networks [7, 10, 13] and calculate configurations in a cooperative fashion on all devices, as relying on central instances is not feasible there. While this approach increases the robustness of the configuration process, it implies extensive communication between the devices. Moreover, it disseminates the configuration tasks equally among all devices, not exploiting resource-rich devices in heterogeneous environments.

In such scenarios, centralized configuration on the fastest device can speed up the configuration process, as it exploits the additional computation power and avoids network communication. As a typical example, [29] presents an efficient centralized approach for weakly heterogeneous Smart Environments, featuring exactly one resource-rich device and several resource-weak devices. To distinguish between the different devices and classify them, [29] uses a combined-metrics clustering strategy to establish a cluster structure consisting of one *cluster head* – the single resource-rich device – which is responsible for centralized configuration calculations, while all other devices are the *cluster members* that remain inactive during configurations. As the cluster head needs to

acquire knowledge of the currently available resources on its cluster members, the *Virtual Container* concept is introduced: A Virtual Container (VC) is a local representation of a remote device on the cluster head and contains the resource information that is relevant for configuration. This enables local access to the remote configuration logic for the cluster head, allowing a strict decoupling of the (re-)configuration processes from the real devices. As the available resources of devices may change over time, each device automatically notifies the cluster head about changes in its resource condition, which then updates the corresponding VC to keep the resource information consistent. After successful configuration, the component bindings that are based on the application structure are established between parent and child components. For runtime adaptation of a configuration, it is sufficient to recalculate only those parts of the configuration that require changes, and re-establish the respective bindings. This scheme represents an efficient solution for weakly heterogeneous Smart Environments. However, such a centralized approach introduces a single point of failure and prevents the parallel calculation of configurations. Furthermore, the other devices' resource information has to be transferred to the configuration device in advance to enable efficient configuration on the resource-rich device. Centralized and decentralized approaches are compared in more detail in [19].

Many typical real-world pervasive scenarios are highly heterogeneous: They feature *several* resource-rich infrastructure devices like servers or desktop PCs as well as small mobile devices such as smart phones or PDAs, like in the auditorium scenario presented in Figure 1. For such environments, we propose a *hybrid configuration approach* in this paper. This approach represents a generalization of the existing centralized and decentralized approaches. It relies on a clustering scheme and enables the application configuration to be computed by multiple resource-rich devices simultaneously, which eliminates the single point of failure that is common in centralized approaches. The resource-weak devices stay passive during the hybrid configuration process. Thus, computational bottlenecks within the calculations are avoided, giving our approach an advantage over fully decentralized approaches. Moreover, our extended clustering mechanism allows the clusters to compute compositions independently from other clusters in the environment. Hence, this hybrid approach reduces the configuration latencies by more than 30% in heterogeneous environments, as our evaluations show.

This paper is structured as follows: After discussing related projects in the next section, we present our system model in Section 3. Afterwards, we introduce our efficient hybrid configuration approach, which is the main contribution of this paper. Then, we present our evaluation results in Section 5 to show the viability of our approach. Section 6 concludes and gives an outlook on future work.

2 Related Work

2.1 Application Configuration and Service Composition

Many current projects deal with component systems for Pervasive Computing. Speakeasy [9] and OSCAR [20] represent exemplary systems allowing users to

create compositions of devices, media and services based on their current context. Pering et al. [24] present a composition framework to enable user-centric collections that combine mobile components together for carrying out a user task. However, these projects rely on user interaction during configuration processes and do not provide algorithms for the automatic composition of application configurations, which is the main focus here.

Projects like Gaia [26], Aura [31], iRoom [17], or Matilda's Smart House [18] provide a middleware for automatic configuration in Smart Environments. They support developers by providing services for the development of context-aware mobile distributed applications. These systems represent highly integrated environments and support various stationary and mobile devices. However, they are not suited for the use in pure ad hoc environments, as they rely on an existing infrastructure. For environments with a higher degree of dynamics, a more recent version of Gaia called Olympus [25] was presented that uses semantic descriptions to automate the mapping process.

In contrast, projects such as Mobile Gaia [7], RUNES [8] or P2PComp [10] target at pure ad hoc networks. While these projects provide automatic configuration, other peer-to-peer based approaches assign this task to the application programmer (e.g., one.world [12]). For highly dynamic environments, Paluska et al. [23] present an indirect specification via goals to refrain from specifying a single configuration. They provide an extensible mechanism to manage users' system runtime decisions and scan the vicinity for techniques that satisfy the user's goals. All of these projects do not rely on a supporting infrastructure, but they also do not exploit the increased computation power of resource-rich devices, yielding suboptimal efficiency in Smart Environments.

MobiGo [30] and PCOM [4] represent systems that support efficient automatic configuration in various environments. While PCOM provides decentralized [13] and centralized [29] configuration algorithms for complex component-based applications, MobiGo focuses on service level virtualization and migration.

Standard component systems like CORBA [21] or Enterprise Java Beans [32] offer persistency and transactional behavior. However, they rather focus on enterprise software than on resource-constrained dynamic pervasive environments. Infrastructures such as Jini [2] or UPnP [16] deal with service discovery in spontaneous networks. Though, they do not provide system support for automatic application configuration and adaptation, which is required here.

Hybrid configuration approaches have already shown to be efficient in other research areas like the distribution of scientific dataflows [3], Web Service composition [5] or large-scale Grid computing systems [11]. With the hybrid scheme presented in this paper, the complete spectrum of possible pervasive environments is covered, giving this system a distinct advance over the related projects that focus on application configuration in specific pervasive environments only.

2.2 Load-Balancing Clustering Schemes

Many related approaches aim at balancing the load among nodes. MANET-based schemes [34] like DEEC [27] or DLBC [1] balance the load in infrastructure-less

scenarios to extend the overall network lifetime. Thus, these schemes equally distribute the load among *all* nodes. In addition, schemes like AMC [22] focus on highly dynamic mobile devices and multi-hop connections. Thus, the merging and split-up of clusters are common actions, yielding low cluster stability. In contrast, we only want to balance the load between the subset of resource-rich infrastructure devices to minimize the (re-)configuration latencies, with as few re-clustering processes as possible. As this infrastructure is typically continuously available, the respective subset of resource-rich devices is rather static. In the area of web clusters, scheduling algorithms try to balance the load distribution on the servers to increase the loading capacity of the cluster [6]. However, these schemes do not consider aspects like mobility or node failures. Hence, they do not provide the re-clustering strategies needed here and are not suited to solve our problem of balancing the configuration load between the resource-rich devices.

3 System Model

3.1 Application Model

For this work, we presume a component-based software model, i.e. an application consists of several *components* which are resident on specific devices and require a certain amount of resources. An application is represented by a tree of interdependent components that is constructed by recursively starting the components required by the root instance. Interdependencies between components as well as resource requirements are described by directed *contracts* which specify the functionality required by the parent component and provided by the child component. A parent component may have an arbitrary number of dependencies. Further details can be found in [4].

3.2 Underlying System

We especially focus on heterogeneous pervasive environments in this paper, consisting of resource-rich devices like PCs or laptops as well as resource-poor mobile devices like smart phones or PDAs. The number of components per device is not restricted. Devices have a unique system identifier (SID) and may become unavailable at any time, e.g., due to mobility or device failures, causing the unavailability of their components. All devices use standard wireless communication technology, e.g., Bluetooth or WiFi, and have a direct, bidirectional communication link to each other, which is usually the case in typical pervasive scenarios like offices or home entertainment. Furthermore, the underlying middleware is supposed to maintain a registry containing all devices in the vicinity with information about their services and properties.

3.3 Problem Statement

We focus on automatic configuration and adaptation of distributed applications in heterogeneous pervasive scenarios. In a *configuration* process, a specific configuration algorithm tries to resolve all application dependencies by finding a

suitable composition of components. Such a composition is subject to two classes of constraints: *Structural constraints* describe what constitutes a valid composition in terms of functionalities. *Resource constraints* are a result of the limited resources. For example, in the presentation application introduced in Figure 1, a structural constraint is that the video projector can only be used if the computer to which the projector is connected to is also available. A resource constraint could be that there needs to be at least one loudspeaker as acoustic output device. An application is successfully configured if all dependencies were resolved and the bindings between the components were established. The *configuration latency* comprises the time between the start and the availability of the application to the user. This latency includes the delays caused by calculating a valid configuration and instantiating all application components. Our goal is to minimize the configuration latency in order to provide a seamless user experience.

Re-configuration processes, or *adaptations*, become necessary if devices whose components are part of the current application configuration become unavailable. Then, alternative components have to be found that can provide the same functionality. Generally, an adaptation represents a special case of a configuration where only those parts of the application need to be recalculated that are no longer valid. So, the same algorithms are used for configuration and adaptation.

4 Hybrid Configuration Management

4.1 Approach and Challenges

Both the totally decentralized and totally centralized approaches have advantages, but also drawbacks that prevent an efficient configuration in *all* possible pervasive environments. Our hybrid approach combines the best properties of these two approaches to minimize the configuration latency. For this purpose, only the resource-rich devices actively calculate application configurations. We call these devices *Active Devices (ADs)* in the following. Contrary to this, the resource-weak devices only provide information about their available resources and services, prior to configuration processes. They stay passive during the configuration, so we call them *Passive Devices (PDs)*.

In a hybrid configuration process, initially, the AD and the PD roles need to be assigned to the devices in the environment since the configuration of each PD has to be calculated by one AD. We call this assignment of a PD to an AD a *mapping*. Subsequently, the ADs need to obtain the configuration-specific information from their mapped PDs. Finally, a hybrid configuration algorithm is necessary which calculates valid configurations on the ADs and distributes the configuration results to the PDs. Details are presented below.

4.2 Cluster Formation and Maintenance

Initially, a suitable subset of devices for calculating configurations has to be discovered to exploit the device heterogeneity efficiently. To reduce the risk of possible bottlenecks, the component algorithms' configuration load should

be balanced between the ADs and maintained even in case of changing device availabilities. Furthermore, each AD should not be required to know about the mappings at the other ADs.

Resource-Aware Cluster Formation. The following scheme establishes multiple stable clusters in heterogeneous environments with several resource-rich devices. These devices automatically become the cluster heads (and, hence, the ADs) if a resource-aware clustering strategy like in [29] is used. Our new scheme balances the configuration algorithm's load among these ADs such that a) they are not overloaded and b) the configuration is parallelized to reduce the latencies.

We assume there are m ADs A_i with cluster indices (CIDs) $i \in \{0, \dots, m-1\}$ and n PDs P_j with indices $j \in \{0, \dots, n-1\}$. Initially, each AD assigns itself a CID i according to its SID, i.e. the AD with lowest SID (of all ADs) assigns itself CID $i = 0$, and the AD with highest SID gets CID $i = m - 1$. The same holds for the PDs that assign themselves CIDs j according to their SID.

There is an overhead for each AD consisting of the efforts needed to retrieve its mapped PDs' resource information, calculate its mapped PDs' components' configuration and send the configuration results back to them. This overhead highly depends on the number of PDs within its cluster. Thus, if the mapping of PDs to ADs is balanced, each AD takes the responsibility for about the same amount of configuration work. This establishes the load balance among the ADs that is important to reduce the configuration latency. To achieve this, each AD has to map at least $\lfloor \frac{n}{m} \rfloor$ PDs to itself. If $n \bmod m = z > 0$, the ADs $0, \dots, z-1$ need to map one additional PD to ensure all PDs are mapped to an AD. This leads to the so-called *Balancing Condition* that has to be fulfilled at each AD:

$$\text{mapped}(A_i) = \begin{cases} \lfloor \frac{n}{m} \rfloor + 1, & i < n \bmod m \\ \lfloor \frac{n}{m} \rfloor, & i \geq n \bmod m \end{cases}, \quad (1)$$

where $\text{mapped}(A_i)$ is the number of PDs that need to be mapped to AD A_i . The fulfillment of this condition is verified on each AD, initially on startup of the device and whenever the number of ADs or PDs changes. For the actual mapping, a simple round robin scheme is used where each AD maps every m -th PD, starting with A_0 that maps P_0, P_m, P_{2m} , and so on.

A mapping procedure is initiated by an AD by sending a mapping request to the PD it wants to map. The PD reacts by transmitting its current resource information to the respective AD so that the AD can create a local representation of the remote PD. This scheme is performed in parallel on all ADs, as they map disjoint sets of PDs. They just need to know their own CID i and the number of ADs and PDs, which can be looked up in the device registry.

For clarification, let us consider an exemplary scenario consisting of three ADs A_0 to A_2 and eight unmapped PDs P_0 to P_7 . Using the described cluster formation scheme, A_0 maps P_0, P_3 , and P_6 . Furthermore, A_1 maps P_1, P_4 , and P_7 , and A_2 maps P_2 and P_5 . The arising cluster structure is shown in Figure 2a.

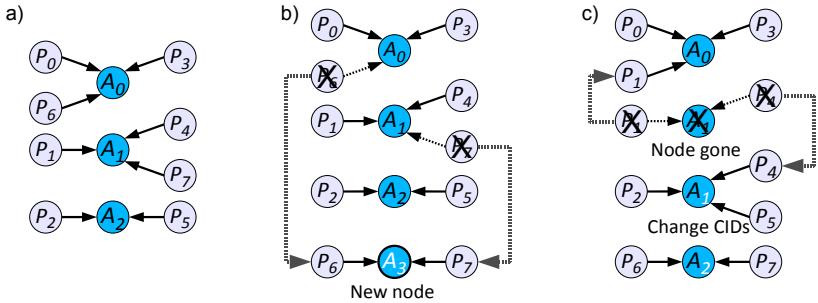


Fig. 2. a) Initial mapping, b) Remapping: A_3 appeared, c) Remapping: A_1 disappeared

Cluster Maintenance. Re-clustering is needed to maintain a balanced load in dynamic environments. Our scheme avoids unnecessary merging and splitting of clusters by simply re-mapping single PDs. Re-clustering comprises four cases: The appearance of a new PD or a new AD, and the disappearance of a PD or an AD. If a new device appears, it assigns itself the lowest free CID within its class, e.g., if it is an AD and there are m other ADs with indices $0, \dots, m-1$ present, it assigns itself index $i = m$. Each device decrements its CID if another device from the same class (i.e., AD or PD) with a lower CID disappears.

If a new PD appears, the AD with cluster index $i = (n \text{ modulo } m)$ maps this device. Thus, the round robin distribution of the PDs to the ADs is continued. After this mapping, each device increments the number n of PDs.

If a new AD D appears, D needs to map $\lfloor \frac{n}{m} \rfloor$ devices to itself¹ so that all ADs still have a similar fraction of PDs. The re-mappings are executed in the following way: Initially, D needs to remap a PD from the AD that has the maximum number of mapped PDs (and the highest index i , in case of multiple options), which yields the AD with index $i = [(n-1) \text{ modulo } (m-1)]$ due to the round robin scheme. D sends a remapping request to the corresponding AD which then notifies its mapped PD with highest CID that this PD has to be remapped to D . This remapping process is repeated $\lfloor \frac{n}{m} \rfloor$ times, whereas the ADs whose PDs are re-mapped by D are chosen by a round robin scheme, as shown in line 5 of Listing 1.1. Hence, the Balancing Condition is still fulfilled on all ADs after these re-mappings. Nodes that appear during an ongoing configuration process – ADs as well as PDs – are not considered within this configuration yet, but starting with the next one. Consider the example from Figure 2b where AD A_3 appeared. At first, A_3 calculates it needs to remap $\lfloor \frac{8}{4} \rfloor = 2$ PDs. According to line 5 from Listing 1.1, A_3 finds out it needs to remap one PD from device $(8-1-0) \text{ modulo } 3 = 1$, and one PD from device $(8-1-1) \text{ modulo } 3 = 0$. Then, A_3 sends remapping requests to these ADs. Thus, A_1 notifies P_7 (its mapped PD with highest index) to remap to A_3 , and A_0 notifies P_6 to remap to A_3 . Now, the Balancing Condition is fulfilled again. If multiple devices appear at almost the same time, the problem of race conditions during the mapping process may

¹ Here, D is already included in the number m of ADs.

arise and potentially lead to inconsistent mappings. To analyze the seriousness of this problem, we did multiple real-world tests where we started two devices timely close to each other and regarded the arising mappings. We found out that inconsistent mappings started to emerge when the time span between two subsequent appearances of new devices fell below 30 ms. Thus, every new device waits for 50 ms for potential other new devices before it starts its mapping process. Proceeding like this, inconsistencies did not appear anymore.

```

1 request_remappings(){
2     mapped := 0;
3     remappings := floor(n/m);
4     while (mapped < remappings) {
5         remapId = (n-1-mapped) modulo (m-1);
6         send_remap_request_to_AD(remapId);
7         await_resource_info();
8         mapped++;
9     }
10 }
```

Listing 1.1. Reclustering process executed by a newly appearing AD

If a PD P_j disappears, all ADs need to decrement the number n of PDs, and P_j 's mapping needs to be removed at the AD A_j to which it was mapped. Additionally, A_j verifies if the Balancing Condition is still fulfilled. If this is not the case, A_j sends a remapping request to the AD with index $k = n \bmod m$. Then, A_k notifies its mapped PD with highest CID that this PD needs to be remapped to A_j . The chosen PD finishes this remapping by sending its resource information to A_j . Additionally, if P_j disappears during an ongoing configuration process, A_j recognizes those parts of the application which were provided by P_j 's components and selects alternative components for them, if available.

Finally, the case of a disappearing AD A_x remains. If A_x was the last available AD, then each PD notices that the cluster structure is dissolved, and the decentralized configuration approach is chosen in future configuration processes. Otherwise, remapping processes are necessary: Each PD that recognizes that its cluster head A_x is gone broadcasts a so-called *Unmapped Message* to notify the other nodes that it is currently unmapped and needs to be remapped to another AD. If an AD A_y notices the disappearance of A_x , it at first checks if A_x had a lower CID than itself. In this case, it decrements its CID. Then, A_y needs to calculate the number of required remappings ($remap(A_x)$) for itself: In order to fulfill the Balancing Condition, A_y needs to remap at least $\lfloor \frac{n}{m} \rfloor$ devices, minus the number of its currently mapped devices ($mapped(A_y)$). As before, if A_y recognizes that there are some remaining unmapped devices, i.e., $n \bmod m = z > 0$, the ADs with indices $0, \dots, z-1$ need to remap one additional device. Subsequently, each AD broadcasts how many remappings it will perform. Then, each AD waits for a certain time T_1 to gather all remapping and unmapped messages. The value of T_1 has to be large enough to cover

the whole gathering process. Otherwise, T_1 expires without all messages having been received, causing inconsistencies and potentially thrashing effects in the remapping processes. However, as too large values of T_1 unnecessarily increase the time for (re-)clusterings, T_1 must also not be chosen too high. A reasonable compromise is to determine the average time a gathering process takes in typical scenarios, and add some additional time to be on the safe side. Further information about the value we chose for T_1 is given in Section 5.1. After this waiting time, each AD knows which PDs are unmapped, and how many PDs the other ADs will remap. Finally, the remappings are performed according to the indices of the involved ADs and PDs: AD A_0 with lowest CID 0 maps the $remap(A_0)$ unmapped PDs with lowest CIDs, i.e. the unmapped PDs with CIDs $0, \dots, remap(A_0) - 1$. AD A_1 with the second lowest CID maps the $remap(A_1)$ PDs with next higher indices, i.e. $remap(A_0), \dots, remap(A_0) + remap(A_1) - 1$, and so on up to the AD with highest CID which maps the unmapped PDs with highest CIDs. In the special case of a disappearing AD A_x during an ongoing configuration process, those parts of the application which were calculated by A_x are no longer available, making a remapping of the PDs that were mapped to A_x and a subsequent restart of the configuration process inevitable. This increases the arising latencies. However, a disappearing infrastructure-device exactly at a configuration process is quite unlikely and should happen rather seldom.

As an example, consider Figure 2: where AD A_1 disappeared, leaving P_1 and P_4 unmapped. Now, A_2 and A_3 decrement their CIDs and become A_1 and A_2 , as an AD with lower CID disappeared. According to the previously described scheme, A_0 and A_1 need to remap one additional device because of their low indices, i.e. $\lfloor \frac{8}{3} \rfloor - 2 + 1 = 1$ PD, while A_2 needs to remap $\lfloor \frac{8}{3} \rfloor - 2 = 0$ PDs. As A_0 has a lower CID than A_1 , A_0 remaps the unmapped PD with lower CID (i.e., P_1), and A_1 remaps P_4 as the unmapped PD with higher CID. Again, the Balancing Condition is fulfilled after these remapping processes.

4.3 Hybrid Application Configuration and Result Distribution

The hybrid configuration is calculated in a parallel and cooperative fashion on the subset of ADs. The configuration of each PD's components is performed locally on the AD it was mapped to. Therefore, the created VCs are used (cf. Section II). This reduces the communication overhead during the configuration compared to decentralized configuration. Moreover, the PDs are not involved in these calculations. This avoids that the resource-constrained PDs become computational bottlenecks, and it conserves their (usually limited) energy resources.

An adapted version [13] of Asynchronous Backtracking [33] is used for the cooperative configuration on the ADs. This decentralized configuration algorithm enables the concurrent configuration of components and utilizes the available parallelism. It performs a depth-first search in the tree of dependencies. In case a dependency cannot be fulfilled, dependency-directed backtracking is used. Furthermore, for the local configuration of the PDs' components on the ADs, we use an efficient centralized algorithm called Direct Backtracking [28]. This algorithm features a proactive mechanism to avoid backtracking in many situations,

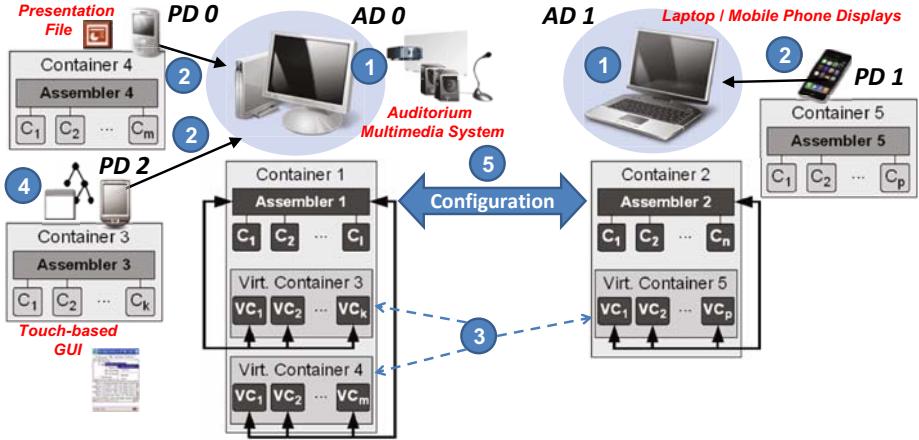


Fig. 3. Hybrid application configuration example

and an intelligent backtracking mechanism to handle conflict situations more efficiently. After a successful configuration, the ADs distribute the configuration results among their PDs to notify them about which of their components were chosen. The respective messages are rather small, as they only contain the relevant information about the chosen components on the recipient PDs. The average message overhead per application component is only 9 kB. Finally, the component bindings are established, yielding the application execution.

4.4 Exemplary Configuration Process

Resuming the scenario from the introduction, Figure 3 shows an exemplary environment where a distributed presentation application needs to be executed. When a speaker wants to give a presentation, the configuration algorithm needs to automatically find suitable components for the distributed application on these devices. For instance, if a speaker wants to switch between the slides using the touchscreen of his mobile phone (PD 2), a touch-based graphical user interface needs to be provided on this device. Moreover, all presentation files may potentially be resident on a remote device, like the conference organizer's smart phone (PD 0). The speaker also needs supporting input and output devices such as the auditorium's multimedia system covering video projector, loudspeakers and a microphone, which are connected to the stationary PC (AD 0). As some auditors may potentially be sitting far from the presentation screen, it might be more convenient for them to have the slides displayed on their own mobile devices, e.g. their laptop (AD 1) or even their mobile phone (PD 1).

Initially, the cluster structure is established using the presented round robin scheme. This yields the desktop PC as cluster head for the PDs 0 and 2, and the laptop as cluster head for PD 1 (step 1). In step 2, the PDs transfer their current resource information to their respective ADs. On the basis of this information, the ADs build the local representations of the mapped PDs within

Virtual Containers in step 3. This leads to two VCs at AD 0, and one VC at AD 1. Then, a user wants to start an application on his/her mobile device, PD 2. Thus, the information about the application start is transmitted to AD 0 as the responsible cluster head for PD 2 (step 4). Subsequently, AD 0 initiates the configuration of the application, which is shown in step 5. At first, it verifies which of the dependencies can be resolved by components of its local container and the Virtual Containers, representing its mapped PDs. For the remaining unresolved functionalities, AD 0 requests AD 1 to resolve these dependencies. AD 1 provides AD 0 with the corresponding information about the fitting components. Subsequently, the complete configuration is constructed by AD 0. After successful configuration, the PDs whose components are used in the configuration are informed by their cluster head about their component configurations. Finally, the required components are initialized, the bindings between the components – as negotiated within step 5 – are established, and the application is executed.

5 Evaluation

5.1 Experimental Setup

For our real-world evaluations, we used six laptops² and six smart phones³. The laptops became cluster heads (ADs) because of their high computation power, while the smart phones became cluster members (PDs) and were equally distributed among the cluster heads. In all scenarios, we used the 802.11b Ad Hoc mode in combination with broadcast messages between the devices. The configuration process was initiated by invoking the application anchor on one of the smart phones. Apart from the real-world experiments, we also performed extensive evaluations on the Network Emulation Testbed (NET, [15]) to evaluate the scalability of our approach in larger scenarios with up to 85 devices. In these evaluations, we emulated the same wireless network as in the real-world evaluations. To find a suitable value for the parameter T_1 for gathering the *unmapped* and *remapping* messages (cf. Section 4.2), we performed 50 measurements to identify the time it takes to gather this information from the other devices. The average time to receive all of these messages was 0.57 s. Furthermore, the gathering process never took longer than 0.83 s, even in large scenarios. As a precaution, we initialized T_1 with a slightly increased value of 1 s for the evaluations. Consequently, we did not face any thrashing effects or race conditions in the remapping processes during any of the taken evaluations. In the shown graphs, each measurement represents the average of 50 evaluation runs. Standard deviations were below 15 % in all cases and below 10% in 90% of all cases.

We used the PCOM [4] system for our evaluations. The evaluated application represents a binary tree of depth 6, i.e., it consists of $k = 127$ components. Additionally, we measured the configuration latencies in a smaller scenario with a binary tree of depth 4, i.e. $k = 31$, to verify our results in a smaller scale. In the

² ThinkPad T41p, Intel Centrino CPU, 1.6 GHz, 1 GB RAM.

³ T-Mobile MDA, PXA 270 CPU, 520 MHz, 128 MB RAM.

evaluations, the laptops got an increased number of resources compared to the smart phones (factor 2 to 5, randomly chosen for each laptop) to consider that they are usually much more resource-rich. We evaluated the hybrid scheme in comparison to the totally decentralized and centralized approaches to show the advantage over these standard approaches. We measured the message overhead and the latencies that arose at the various stages of the configuration: initial cluster formation and re-clustering processes, the preconfiguration process, the actual configuration as well as an adaptation process where only 50 % of the components needed to be adapted, the distribution of the configuration results, and the binding of the components.

5.2 Communication Overhead Measurements

Figure 4 shows the message overhead at the various stages of the configuration. In these graphs, “Hybrid- x ” represents the hybrid approach with x ADs (laptops), where $2 \leq x \leq 6$. The remaining devices (PDs) were the smart phones.

In the preconfiguration process (Figure 4a), an average overhead of 53 kB per device and configuration process arises for the centralized and hybrid schemes, since these schemes need to build the cluster structure and to transmit the configuration-specific information for the VCs. For hybrid configuration, this overhead arises only at every PD, as they need to transmit their resource information to their cluster head. This leads to a reduced overhead compared to the centralized scheme. The decentralized scheme does not use preconfiguration.

Figure 4b shows the message overhead for the actual configuration. In centralized configuration, the device where the application was started initially transmits the application information to the cluster head. The resulting overhead only depends on the application size, i.e. the involved components. As we used a fixed application with 127 components, the overhead was static with 183 kB in total per configuration process. The hybrid approach’s message overhead mainly depends on the number of involved ADs, as only they calculate configurations. Thus, a rising number of available PDs does not have an impact on the message overhead. The message overhead for decentralized configuration increases with a rising number of involved devices, as all devices have to communicate with each other. However, this overhead converges for a larger number of involved devices, since the per-device-overhead decreases due to a lower number of components per device. The centralized approach’s distribution overhead (Figure 4c) and the component binding overhead (Figure 4d) converge for the same reason.

As the devices piggyback the configuration results during the decentralized configuration process, no further messages are needed for result distribution, as it can be seen in Figure 4c. Compared to the centralized approach, the piggy-backing increased the overhead during the actual configuration by 403 kB, but reduced the result distribution overhead by 1418 kB on average. In centralized configuration, the cluster head broadcasts the *complete* composition, yielding high communication overhead. In hybrid configuration, the cluster heads only need to notify their PDs about which of their components were chosen. Thus, the hybrid approach’s overhead rises linearly with the number of PDs.

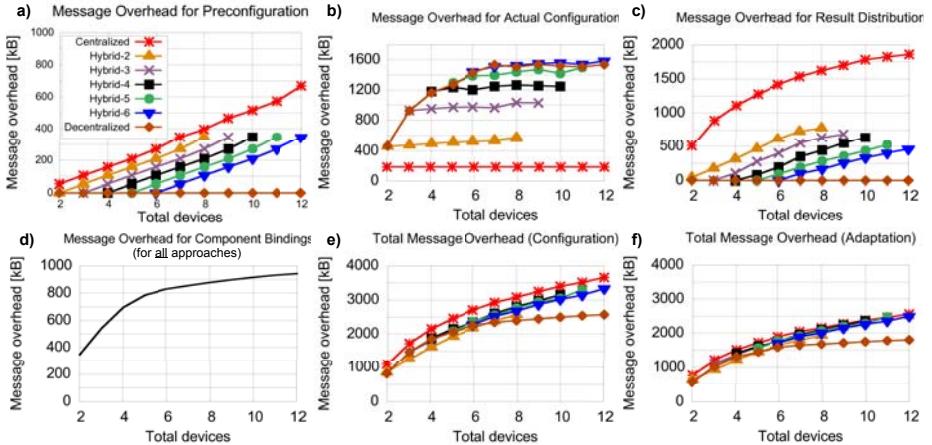


Fig. 4. Message overhead at the different stages of one configuration process ($k = 127$)

The overhead for establishing the component bindings (Figure 4d) is the same for all configuration schemes, as it is independent from the actual configuration. This overhead rises with a rising number of involved devices, since bindings between components on different devices are likely to emerge more often then.

Figure 4e shows the total message overhead for one configuration process as the sum of all overheads. The decentralized approach scales best due to the result piggybacking at the configuration process. Its total message overhead converges with a rising number of involved devices due to the almost constant overhead for actual configuration and no further distribution overhead (cf. Figures 4b and 4c). The centralized approach performs worst because of a high overhead for preconfiguration and result distribution. The hybrid approach produces an average overhead at all stages of configuration, yielding a moderate total overhead and showing its applicability concerning message overhead.

Regarding adaptation, the total message overhead is shown in Figure 4f. Compared to configuration, the overheads for the centralized and decentralized schemes were reduced by 30 %, as only parts of the application needed to be recalculated and distributed. The message overhead of the hybrid scheme decreased by 25 % only, as the remapping messages needed to be sent, too. Thus, the hybrid and centralized schemes produce about the same adaptation message overhead, while the decentralized schemes' overhead is around 22 % lower.

5.3 Configuration Latency Measurements

We compared the overall latencies of all three approaches in two heterogeneous scenarios ($k = 31$, $k = 127$) with differing device numbers and 50 % resource-rich devices in each scenario. Figure 5 shows the total latencies. The real-world evaluations were performed with 4 to 12 devices, and the emulations in the large-scale scenario with $k = 127$ with up to 85 devices, where each laptop holds two

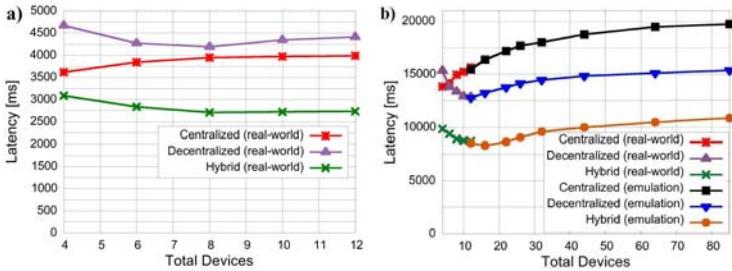


Fig. 5. Total configuration latencies: a) $k = 31$, b) $k = 127$

resources and each smart phone holds one resource. Increasing the number of devices above 85 would not lead to changing results, since some of the devices would not hold any resources then. Figure 5b shows that the latencies for the hybrid and the decentralized approach at first drop with a rising number of devices. This happens because of an increasing *absolute* number of resource-rich devices that are involved in configuration calculations, while in centralized configuration, *only one* resource-rich device is always used to calculate configurations. When the total number of devices exceeds 12 (distributed) or 16 (hybrid) devices, the overall latencies start to slightly increase again, as the latencies for establishing the component bindings grow stronger than the latencies for the configuration calculation drop. The latencies of centralized configuration show continuous growth, as the latencies for distribution and establishment of the bindings increase with a rising number of devices, while the configuration latency remains constant. It can be seen that the hybrid approach outperforms the decentralized approach by 35.7 % ($k = 31$) and by 34.5 % ($k = 127$) on average, and the centralized approach by 26.3 % ($k = 31$) and by 44.1 % ($k = 127$), respectively. The emulation results point up the hybrid approach's scalability, as latency reduction still holds with large applications and many involved devices.

For clarification, Figure 6 shows the latencies at the different configuration stages in a specific scenario with $k = 127$, four ADs and up to six PDs. The clustering of devices produces a negligible latency of below 30 ms per PD, as you can see in Figure 6a. Re-clustering processes due to dynamics take a constant time of 1.1 s more than the initial clustering, mainly because of the chosen value of 1 s for T_1 (cf. Section 5.1). The loading of the resource information increases linear with an overhead of 400 ms per device. The clustering and resource information loading latencies are *not* included in the overall latencies in Figures 6e and 6f, as they are performed once *prior* to the configuration. However, the re-clustering latency is included in the overall adaptation latency shown in Figure 6f.

Regarding the latency for the configuration process itself (Figure 6b), the centralized approach performs best, as the resource-richest device locally calculates the configuration. The decentralized approach is significantly slowed down due to the fact that the resource-limited devices are involved in the calculations. Another factor is the immense communication overhead of the decentralized

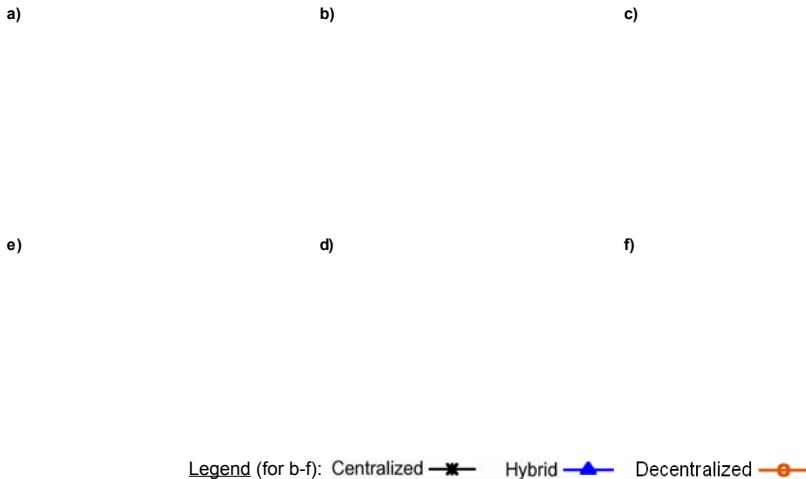


Fig. 6. Latencies at the different stages of the configuration process ($k = 127$)

approach (cf. Figure 4b). In the hybrid approach, only the resource-rich devices perform the calculation, but message exchanges between them still take time. Thus, its latencies are slightly above the centralized scheme's latencies.

Figure 6c shows the latency to distribute the configuration results. The centralized scheme has the highest latency, as the single configuration device needs to distribute the complete configuration (cf. Figure 4c). In contrast, the other approaches have already piggybacked information about configured components in the configuration messages, in case of decentralized configuration even between *all* devices. Thus, these approaches have much lower distribution latencies.

The initialization of the component bindings (Figure 6d) comprises the sum of the import of the received configuration results and the establishment of the respective component links. Since message overhead and delay for the result distribution are much higher for the centralized approach, as seen in Figures 4c and 6c, the configuration import is responsible for a big fraction of the latency, especially on the resource-weak devices. The establishment of the links is performed in the same way by all approaches and, hence, takes the same amount of time.

Figure 6e shows the total latencies as sum of the latencies from Figures 6b-d. The centralized approach is slowest due to its increased result distribution and component binding overhead. The decentralized scheme performs 14 % better on average, although the resource-weak devices are involved. The hybrid approach avoids the drawbacks of the other schemes and performs fine in all configuration stages. Thus, it outperforms the decentralized scheme by 34.2 % and the centralized scheme even by 40.7 % on average. Regarding the total latencies for an adaptation process (Figure 6f), the advantage of the hybrid approach decreases to 20.4 % compared to decentralized and to 30.2 % compared to the centralized scheme, due to the additional re-clustering overhead (cf. Figure 6a).

6 Conclusions and Outlook

We presented a hybrid approach for configuring distributed pervasive applications. This approach efficiently exploits the available computation resources in heterogeneous environments. Since this hybrid scheme is a generalization of the pure centralized and decentralized approaches, it covers the complete spectrum of pervasive scenarios, which has not been achieved by related projects yet.

Our approach is based on the formation of clusters with balanced configuration load for the resource-rich devices. These devices represent the active devices during configuration calculation processes, while the resource-weak devices remain passive to avoid bottlenecks in the configuration process. Single points of failure are avoided due to the parallel execution of the configuration calculations on the active devices. The hybrid approach automatically adjusts its degree of decentralization to the available resources in the network. In our evaluations, we proved that our approach reduces the configuration latencies by more than 30 % on average compared to decentralized and centralized approaches. Moreover, the evaluations on a network emulation cluster showed that these results also hold in larger scenarios. The reduced configuration time strongly helps to increase users' acceptance for pervasive systems and represents a large step towards seamless automatic application configuration.

Our next step is to reduce the hybrid approach's communication latencies by exploiting the application structure and local component dependencies at the clustering processes. Moreover, we want to use idle periods at the ADs to precalculate partial configurations and store them for future configuration processes.

References

1. Amis, A.D., Prakash, R.: Load-Balancing Clusters in Wireless Ad Hoc Networks. In: Proc. IEEE Asset 2000 (2000)
2. Arnold, K., O'Sullivan, B., Scheifler, R., Waldo, J., Wollrath, A.: The Jini Specification. Addison Wesley, Reading (1999)
3. Barker, A., Weissman, J.B., van Hemert, J.: Eliminating The Middleman: Peer-to-Peer Dataflow. In: Proc. ACM HPDC 2008 (2008)
4. Becker, C., Handte, M., Schiele, G., Rothermel, K.: PCOM - A Component System for Pervasive Computing. In: Proc. IEEE PerCom 2004 (2004)
5. Benatallah, B., Sheng, Q.Z., Dumas, M.: The Self-Serv Environment for Web Services Composition. IEEE Internet Computing 7(1) (2003)
6. Cardellini, V., Colajanni, M., Yu, P.S.: Dynamic Load Balancing on Web-server Systems. IEEE Internet Computing 3(3) (1999)
7. Chetan, S., Al-Muhtadi, J., Campbell, R., Mickunas, M.D.: Mobile Gaia: A Middleware for Ad-hoc Pervasive Computing. In: Proc. IEEE CCNC 2005 (2005)
8. Costa, P., et al.: The RUNES Middleware for Networked Embedded Systems and its Application in a Disaster Management Scenario. In: Proc. IEEE PerCom 2007 (2007)
9. Edwards, W.K., et al.: Using Speakeasy for Ad Hoc Peer-to-Peer Collaboration. In: Proc. ACM CSCW 2002 (2002)
10. Ferscha, A., Hechinger, M., Mayrhofer, R., Oberhauser, R.: A Light-Weight Component Model for Peer-to-Peer Applications. In: Proc. ICDCS 2004 Workshops (2004)
11. Graupner, S., Andrzejak, A., Kotov, V.E., Trinks, H.: Adaptive Service Placement Algorithms for Autonomous Service Networks. In: Proc. ESOA 2004 (2004)

12. Grimm, R.: One.world: Experiences with a Pervasive Computing Architecture. *IEEE Pervasive Computing* 3(3) (2004)
13. Handte, M., Becker, C., Rothermel, K.: Peer-based Automatic Configuration of Pervasive Applications. In: Proc. IEEE ICPS 2005 (2005)
14. Handte, M., Urbanski, S., Becker, C., Reinhard, P., Engel, M., Smith, M.: 3PC-/MarNET Pervasive Presenter. In: Proc. IEEE PerCom 2006 (2006)
15. Herrscher, D., Rothermel, K.: A Dynamic Network Scenario Emulation Tool. In: Proc. ICCCN 2002 (2002)
16. Jeronimo, M., Weast, J.: UPnP* Design by Example. Intel Press (2003)
17. Johanson, B., Fox, A., Winograd, T.: The Interactive Workspaces Project: Experiences with Ubiquitous Computing Rooms. *IEEE Pervas. Computing* 1(2) (2002)
18. Lee, C., Nordstedt, D., Helal, S.: Enabling Smart Spaces with OSGi. *IEEE Pervasive Computing* 2(3) (2003)
19. Liu, D., Law, K.H., Wiederhold, G.: Analysis of Integration Models for Service Composition. In: Proc. ACM WOSP 2002 (2002)
20. Newman, M.W., Elliott, A., Smith, T.F.: Providing an Integrated User Experience of Networked Media, Devices, and Services Through End-User Composition. In: Indulska, J., Patterson, D.J., Rodden, T., Ott, M. (eds.) PERVASIVE 2008. LNCS, vol. 5013, pp. 213–227. Springer, Heidelberg (2008)
21. Object Management Group (OMG): CORBA Component Model V3.0 (2002)
22. Ohta, T., Inoue, S., Kakuda, Y.: An Adaptive Multihop Clustering Scheme for Highly Mobile Ad Hoc Networks. In: Proc. IEEE ISADS 2003 (2003)
23. Paluska, J.M., Pham, H., Saif, U., Chau, G., Terman, C., Ward, S.: Structured Decomposition of Adaptive Applications. In: Proc. IEEE PerCom 2008 (2008)
24. Pering, T., Want, R., Rosario, B., Sud, S., Lyons, K.: Enabling Pervasive Collaboration with Platform Composition. In: Tokuda, H., Beigl, M., Friday, A., Brush, A.J.B., Tobe, Y. (eds.) Pervasive Computing. LNCS, vol. 5538, pp. 184–201. Springer, Heidelberg (2009)
25. Ranganathan, A., Chetan, S., Al-Muhtadi, J., Campbell, R.H., Mickunas, M.D.: Olympus: A High-Level Programming Model for Pervasive Computing Environments. In: Proc. IEEE PerCom 2005 (2005)
26. Román, M., Hess, C.K., Cerqueira, R., Ranganathan, A., Campbell, R.H., Nahrstedt, K.: Gaia: A Middleware Infrastructure to Enable Active Spaces. *IEEE Pervasive Computing* 1(4) (2002)
27. Safa, H., Mirza, O., Artail, H.: A Dynamic Energy Efficient Clustering Algorithm for MANETs. In: Proc. IEEE WIMOB 2008 (2008)
28. Schuhmann, S., Herrmann, K., Rothermel, K.: Direct Backtracking: An Advanced Adaptation Algorithm for Pervasive Applications. In: Proc. ARCS (2008)
29. Schuhmann, S., Herrmann, K., Rothermel, K.: A Framework for Adapting the Distribution of Automatic Application Configuration. In: Proc. ACM ICPS 2008 (2008)
30. Song, X., Ramachandran, U.: MobiGo: A Middleware for Seamless Mobility. In: Proc. IEEE RTCSA 2007 (2007)
31. Sousa, J.P., Garlan, D.: Aura: an Architectural Framework for User Mobility in Ubiquitous Computing Environments. In: Proc. IEEE/IFIP WICSA (2002)
32. SUN Microsystems: Enterprise Java Beans Specification, Java Specification Request (JSR) 220 Final Release (2003).
<http://java.sun.com/products/ejb/docs.html>
33. Yokoo, M., Durfee, E.H., Ishida, T., Kuwabara, K.: The Distributed Constraint Satisfaction Problem: Formalization and Algorithms. *IEEE Transactions on Knowledge and Data Engineering* 10(5) (1998)
34. Yu, J.Y., Chong, P.H.J.: A Survey of Clustering Schemes for Mobile Ad Hoc Networks. *IEEE Communications Surveys and Tutorials* 7(1) (2005)

12Pixels: Exploring Social Drawing on Mobile Phones

Karl D.D. Willis^{1,2} and Ivan Poupyrev^{2,*}

¹ Carnegie Mellon University

5000 Forbes Avenue, Pittsburgh, PA 15213 USA

² Disney Research Pittsburgh

4615 Forbes Avenue, Pittsburgh, PA 15213 USA

{karl, ivan.poupyrev}@disneyresearch.com

Abstract. In this paper we present the design and development of *12Pixels*, a novel interface, application, and social web service that allows people to create and share drawings directly from a regular mobile phone. We detail the release of *12Pixels* as a service in Japan and analyze trends that emerged from user data collected. Our analysis and insights provide useful ground-level experiences with social drawing and mobile content creation.

Keywords: 12Pixels, twelve pixels, drawing, mobile phone, cellphone, art, design, creativity, creativity support tools, content creation, user generated content, social web, web applications.

1 Introduction

In recent years there has been an explosion of tools and services facilitating everyday creativity. The shrinking size of digital devices such as digital cameras and mobile computers has allowed people to create content at any time and any place. At the same time the proliferation of the internet has supported the distribution of people's creations and lead to a creative boom of 'user generated content' and 'social web' services. The user generated content model brings a ground-up approach to content creation and marks a profound change for computer users and society as a whole.

Despite the growing interest in online content creation, studies have shown that overall time spent with media in the home has not changed significantly for young people since 1999 [1], and has shown overall declines in the three years till 2009 [2]. Any desktop based tool for content creation therefore competes with a large variety of traditional media for an increasingly smaller slice of our free time. However, the ubiquitous availability of modern mobile phones makes them an ideal device to be used for creative means. Current mobile phones are powerful, rich in functionality, and network connected – they have become an important ingredient of everyday culture. Designing interfaces and applications that allow people to create and share content on the mobile phone is an important and crucial direction for pervasive computing technology.

* This project was conducted when both authors were working at Sony Computer Science Laboratories in Tokyo, Japan.

In this paper we present the design and development of *12Pixels*, a novel interface, application and social web service that allows people to create and share drawings directly from their mobile phone. The project began at Sony Computer Science Laboratories in 2007 with the question: '*How can we enable people to better express themselves with mobile phones?*'. We observed that while most standard mobile phones offer rich text and camera capabilities, they lack the fundamental ability to draw and communicate pictorially. The goal of our research was to develop a native drawing interface for the mobile phone (Figure 1).



Fig. 1. Drawing on a mobile phone with *12Pixels*

We chose to design a drawing interface for traditional keypad-based mobile phones. The emerging popularity of touch-screen based mobile phone interfaces, such as those found on *iPhone*, has re-introduced many desktop computer drawing techniques onto mobile devices. However, in spite of growing popularity, such high-end 'smart phones' still account for a surprisingly small segment of global mobile phone users. According to recent reports the *iPhone* is still used by less than 1% of users world-wide [4] and has had particular difficulties in Japan where the *12Pixels* project is based. Therefore, developing tools that allow truly everyone to be creative means developing tools for the traditional mobile phone – a challenging interface problem. The design of a unique interface for drawing on the standard mobile phone is the first important contribution that we report in this paper.

12Pixels was released in Japan to the general public as a social web service that allows drawings to be created, shared, and remixed by a wide range of people using a standard mobile phone. *12Pixels* was the first mobile web service for 'social drawing' and we introduce this system as the second contribution of this paper.

By releasing *12Pixels* as a public service we saw it as an opportunity to evaluate our design approach, learn more about our users, and iteratively evolve the system as

a whole. Would people be interested in drawing on a mobile phone? Would they be prepared to contribute and share their drawings? What type of drawings would they create? When they would create them? What kind of functionality would be desired? Although our designs were informed by a range of user observations, the questions above could only be answered by releasing the application and analyzing usage patterns and content from real users. We present analysis and evaluation of the real-world deployment of *12Pixels* as the third unique contribution of our paper.

The remainder of the paper is organized as follows. In the next section we discuss related work and present a summary of content creation and drawing applications for the mobile phone. We then take a detailed look at the *12Pixels* drawing interface and social drawing service, and explore some of the design decisions behind its development. We next analyze important trends that emerged from user data collected during the release of *12Pixels* in Japan, and reflect upon the success of the project. We believe that our analysis and insights have implications beyond the context of drawing and can inform designers and developers in creating richer and more engaging tools for mobile content creation. Finally we discuss future work and identify the most promising research directions in this area.

2 Related Work

Tools that facilitated content creation on mobile phones were initially very few and very limited in their capabilities. Text based input, available on almost any mobile phone, has been used creatively for some time with everyday communication decorated with text-based icons called ‘emoticons’. In Japan, text-based emoticons are referred to as *kaomoji* (face characters) and progressed from simple character combinations to increasingly complex designs expressing a wide range of emotions (Figure 2).

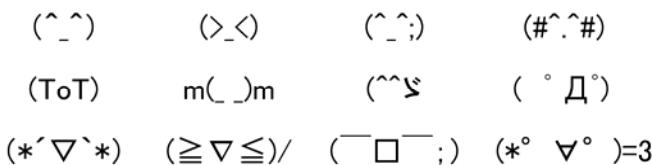


Fig. 2. A large variety of *kaomoji* (text based emoticons) has been invented in Japan by mobile phone users and are actively used in text-based communication

Emoticons have not been the only venue for text-based creative expression on mobile phones. ‘*Keitai shosetsu*’ (mobile phone novels) were born in Japan as a new literary genre; each chapter is written on the mobile phone and then distributed via text message. Mobile phone novels have acquired such popularity that in 2007 four out of five top novels sold in Japan originated on the mobile phone [5]. The exact reasons for the popularity of these novels is debated, however it is worth noting they possess qualities unique to the mobile phone and are not easily replicated on the printed page.

Similar to text, mobile photography has become another popular venue for self-expression. Typical services include photo messaging, photo-based mobile blogging, online photo sharing, and mobile photography competitions. Mobile phone music creation has been the subject of isolated research and design projects and impacted users most in the form of ringtone creation applications. The Digital Minimo D319 handset, released in Japan in 1996, was the first handset that allowed users to compose their own ring-tone melodies. What followed was a brief period of popularity with the release of numerous books listing the ringtone notes of popular songs [6].

Perhaps the earliest attempt at implementing drawing functionality on the standard mobile phone, is the *Draw Picture* application provided on the Sony Ericsson T68 from 2001 onwards. The mobile phone keypad is used to move the pen and draw in a ‘turtle graphics’ style with imagery saved as monotone bitmap files. The *CONTE* (Canvas ON mobile TElephone) drawing tool developed at Shizuoka University [7] in 2002 was the earliest academic exploration into drawing on a standard mobile phone. As well as offering ‘turtle graphics’ style pen controls, *CONTE* also provided more advanced functionality such as stamps, text input, online album upload, and drawing playback. Further work on the *CONTE* system introduced a technique for drawing curved lines using the standard keys of the mobile phone [8]. While both *Draw Picture* and *CONTE* explore interfaces for drawing on the mobile phone platform, they borrow heavily from existing pointer-based interfaces where the user typically creates marks using a pen or mouse. Standard mobile phones do not typically support this form of interaction, making control of the pointer with buttons cumbersome at best. These projects also pre-date the emergence of user generated content and the social web, and therefore do not fully explore the possibilities of community driven content creation.

There have since been numerous approaches to drawing on mobile phones using pen-stylus and touch-screen interaction, however, due to our focus on conventional mobile phone handsets we will not elaborate further on these systems in this paper. An altogether different approach is presented in the *TinyMotion* [9] system by utilising a mobile phone camera and computer vision algorithms to allow users to write letters in the air for text input. While not intended as a drawing system, *TinyMotion* does offer a natural form of gestural interaction that lends itself to drawing. However this form of interaction is not always feasible for mobile phone users in crowded or cramped locations. *TinyMotion* also has very specific hardware and software requirements, meaning it can only be used on a limited subset of mobile phone devices.

Recent years have seen the release of a range of more powerful mobile platforms, which have greatly widened the range of tools available for mobile creativity and content creation. Although much excitement has been generated, it remains to be seen how these new platforms can contribute to mobile phone culture and creativity at large. The *12Pixels* project has been firmly grounded in and inspired by our direct experience with Japanese mobile phone culture. We also drew inspiration from the ever-growing popularity of social tools that encourage people to create, upload, share, remix, reuse, rate, and comment on one another’s creations. Through our research we attempted to instigate a culture of drawing on the mobile phone in a similar vein to the social web, we label this ‘social drawing’. We embarked on the task of transforming the ordinary mobile phone into a social drawing tool, much like a new type of ‘digital brush’, allowing people to draw with very simple input devices, but express themselves and communicate in numerous ways. In the rest of this paper we present what we were able to achieve.

3 Drawing with *12Pixels*

The *12Pixels* drawing interface has gone through numerous iterations since it was first developed in early 2007 [3]. In this section we present the most recent application currently deployed by the *12Pixels* service in Japan, and highlight some of the design decisions made during the iterative design process.

Jenkins observes that *low barriers to artistic expression* are one of the key requirements of what he labels ‘participatory culture’ [10]. Tools for content creation are therefore one of the fundamental ingredients of a user generated content ecosystem. Such tools must not only be easy to use, but should be usable straight away without learning, and able to produce immediate results. In the case of drawing, the user should be able to start drawing on the mobile phone just as easily as they do when using pen and paper.

The importance of immediacy is also stressed in a number of disciplines related to creativity. In developing programming languages for children, Papert uses the metaphor of *low floors* and *high ceilings* [11]. Meaning the language should be easy to get started in but still allow for increasingly complex projects over time. Through their work designing construction kits for children, Resnick and Silverman extend this metaphor by introducing the idea of *wide walls* where a wide range of different projects can be created [12]. These were key considerations in designing the *12Pixels* interface to accommodate a wide audience and a range of different usages.

3.1 12Pixels Basics

The *12Pixels* drawing technique is based on the fundamental idea that each of the twelve keypad keys on the mobile phone can be mapped to one of twelve pixel cells onscreen. Each pixel cell is not a single pixel but rather a section of the drawing area that mirrors the three by four grid of the keypad. As each cell corresponds to a button on the mobile phone keypad, when the user presses the top-left ‘1’ key, its corresponding cell in the top-left of the drawing area is marked out. This spatial relationship between the keypad buttons and the onscreen interface is immediately apparent and understandable for first time users. Figure 3 shows the spatial relationship between keypad keys and pixel cells onscreen.

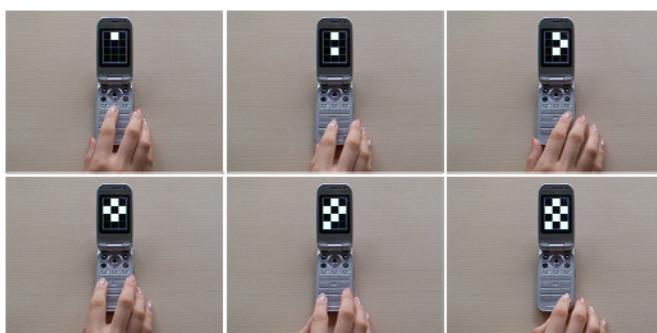


Fig. 3. Drawing with *12Pixels*, the buttons of the standard mobile phone are mapped to pixel cells on screen to create imagery

Similar to other projects within the field of embodied interaction [13] this technique takes advantage of strong tactile cues that the physical mobile phone keypad provides. At any moment the user can find and press relevant keys to mark out corresponding pixels onscreen, relying *only* on tactile feedback i.e. by sliding their fingers over the phone keypad. This is especially important for mobile phone users, as they need to repeatedly shift their attention from the mobile phone to the surrounding environment.

In designing the *12Pixels* interface we also attempted to minimize the number of key presses required when drawing. At any moment the user can select from twelve neighboring pixel cells without having to repeatedly reposition a pointer or cursor. This makes drawing a shape such as a plus sign (+) simply a matter of pressing five different keys on the keypad. Using a pointer-based drawing approach would require an additional six key presses in order to move the pointer then mark each pixel out one by one. The pointer approach also requires significant finger movement back and forward between the keypad and the directional keys.

Our approach also allows users to operate the interface with only *a single thumb*. This is a crucial requirement for any mobile phone interaction as it frees up the other hand for carrying other objects such as a bag or an umbrella, or holding on to a safety railing in the train or on an escalator.

3.2 Drawing Levels

Spatially linking physical buttons to onscreen interface elements has been used to some degree in existing mobile phone menus and also explored as an interaction technique for previewing information [14], zooming [15], and selecting on-screen elements [16]. *12Pixels* expands upon these techniques to implement a drawing interface with ‘levels’ that allow incrementally smaller cells to be marked out on-screen for the creation of more detailed and sophisticated images.

The directional keys of the mobile phone ‘joystick’ and the center selection key are used to control and navigate through levels. As the user presses the selection key, the

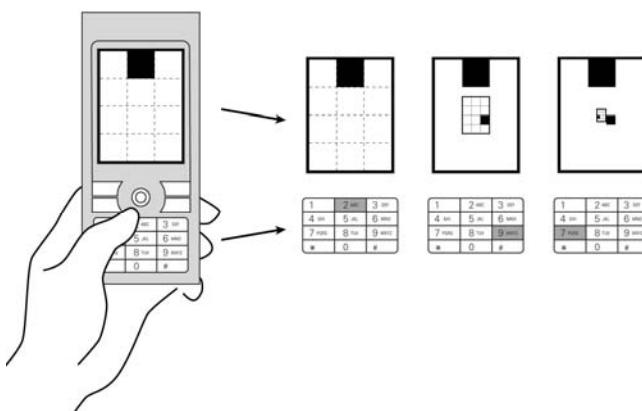


Fig. 4. The directional keys can be used to toggle between levels and move the drawing area

drawing area shrinks down a level and can then be repositioned onscreen with the directional keys. Using the keypad the user can then mark out smaller sized cells onscreen to create imagery at a finer scale. Currently three levels have been implemented, with each level becoming incrementally smaller in a fractal like way, as shown in Figure 4. At the top level the drawing area controls the maximum 27 x 36 cell area, with the middle and lower level controlling 9 x 12 and 3 x 4 cells respectively.

3.3 Additive and Subtractive Drawing

Large areas of the screen can be quickly filled to allow drawing using both additive and subtractive methods. It is just as easy to add pixels to a blank area as it is to subtract pixels from a drawn area. Aside from filling each cell individually, the drawing area itself can also be filled with a single press of the clear key. As illustrated in Figure 5, to draw a simple character the user can quickly fill in large areas for the head and body, then mark out finer details such as the eyes, ears, and mouth by subtracting from the drawn area.

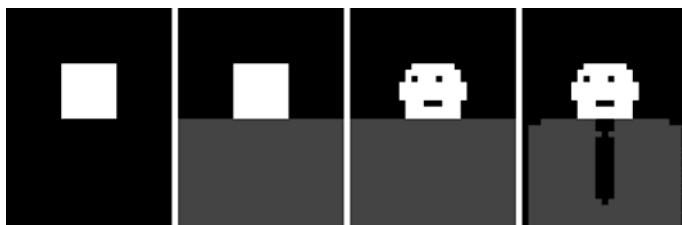


Fig. 5. Additive and subtractive drawing with the *12Pixels* interface

As opposed to traditional drawing, where the artist repeatedly adds color to the canvas, this approach is more akin to sculpting with clay or play-doh. Users can quickly block out key elements before adding further detail. The drawings themselves reflect this process, showing strong composition and contrast.

3.4 Color and Lines

The initial design iteration of the *12Pixels* interface allowed only grayscale drawings to be created. Our motivation was to make the drawing process as simple as possible, akin to sketching pencil lines on paper. However, during the first round of *12Pixels* workshops, color was the single most requested feature. To address this issue we experimented with a number of different designs ideas for adding color in a way that would not impede upon the simplicity of the interface. We explored the use of monotone colors schemes, color blending, and the mobile phone camera to select colors from the physical environment. We decided on using a function key to activate a simple color menu with a preset primary color palette. Shades of color can then be selected by toggling through a sequence of color gradations from dark to light. The color is then stored so that when marking out subsequent pixel cells the current color is maintained.

In addition to marking out pixel cells using single key presses, *lines* can be created by holding down a keypad key then moving the joystick in the direction desired. The drawing area shrinks down to the size of the single pixel cell according to the level, allowing for lines of different thickness to be created.

3.5 Preliminary Observations

In designing *12Pixels* we strived to develop an interface that could be used almost immediately with minimal time spent learning. We wanted users to begin drawing at once, and learn the more advanced features of the interface, such as drawing with shades of color or lines, along the way.

It is often the case that the developers of tools and interfaces become very proficient in their use, however, what would an average person be able to draw with *12Pixels*? To answer this question we conducted a two month long experiment where the application was provided to members of the general public visiting the Sony ExploraScience centre in Tokyo. The range of drawings produced by visitors to the exhibition was surprisingly extensive and showed a wide range of skill for first time users. Looking at the quality of drawings fuelled our enthusiasm for the project and gave us some valuable insights into the type of drawings people want to create.

We next conducted two workshops at the Science Gallery in Dublin. The aim was to gather direct feedback on the drawing experience, establish ways in which the interface could be improved, and test out the initial implementation of our social drawing service. The first group consisted of 13 high school students aged between 15-16 and the second group were 10 adults aged between 20-40. We began with a simple demonstration followed by a period of drawing time where we observed how quickly they came to grips with the interface. A communal screen was setup so that participants could optionally upload their drawings to be displayed to everyone else. This feature was used heavily in both workshops as participants were eager to share their drawings with one another. Much to our surprise it was common for written messages to be created and uploaded; this proved to be a popular way to use *12Pixels* in both the initial workshops and the subsequent public release.

These early user observations were an invaluable experience and proved to us several important points. Firstly, the Sony ExploraScience exhibition showed us that drawings of high quality and wide creative range could be created by first time users. Secondly, the Science Gallery workshops proved to us first hand that sharing one's creations is a key component of the creative process. These points encouraged us to continue improving *12Pixels* and further develop our concept of 'social drawing'.

4 Social Drawing

Jenkins notes that another fundamental element of 'participatory culture' is the infrastructure for sharing content and building a community of like-minded creators [10]. In this section we outline details of the *12Pixels* infrastructure that share many of the aims outlined by Jenkins to support and encourage content creation.

4.1 Release and Implementation Details

12Pixels was released publically as a free service via Sony Style Japan in March 2009. The Japanese market we believed would be an ideal real-world situation to test both the *12Pixels* service and our ideas about mobile content creation and social drawing.

To date there have been two official releases of *12Pixels* in Japan. The first release consisted of monotone drawing with five differently colored applications available in black, red, blue, green, and pink. Users could select their favorite color to download and draw using its five shades. This approach allowed us to test the core interface design in the field, and evolve the use of color in later releases based on user feedback. The second version of *12Pixels* was released in late July 2009 and implemented full color drawing, along with a range of updated web service functionality such as commenting, ranking, and advanced gallery features.

12Pixels is implemented as a Java-based frontend mobile phone application, with a web server backend used to exchange drawings and information from a central database. Three separate applications are compiled and distributed for the main Japanese mobile networks, NTT docomo, Softbank Mobile, and au by KDDI. Despite recent trends towards install-free web applications for desktop computers, this approach has yet to become a reality for mobile phones. By dividing the functionality between a frontend mobile application and a backend server we gain the superior functionality and reliability of an installed application with online connectivity when needed. In the discussion that follows we present the core features of *12Pixels* that facilitate social interaction between users.

4.2 Social Interaction

The *12Pixels* service offers much of the functionality that has become common in desktop based user generated content web applications. It supports the distribution and sharing of drawings both through formal online services and via user-to-user messaging.

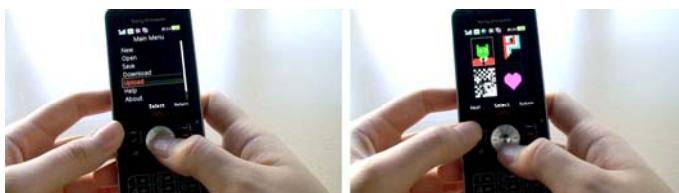


Fig. 6. Drawings can be shared by uploading (left) and downloading (right) directly within the *12Pixels* application

Sharing, browsing, and searching drawings. Drawings created in *12Pixels* can be uploaded with descriptive keywords directly from the mobile phone application. Drawings are then added to the communal gallery on the *12Pixels* server and can be accessed immediately from either the *12Pixels* application or a web browser. From within the *12Pixels* application the drawing gallery can be browsed by date created, keyword, and popularity rating. When a user finds a drawing they like, it can then be

downloaded directly to the *12Pixels* drawing canvas for further editing and remixing (Figure 6). All content submitted to the communal gallery is accessible to all users, and can be freely downloaded and edited at any time.

Commenting and rating. Comments can be added to drawings by browsing to the drawing inside the application and then launching the drawing's comment page in the mobile web browser. This page displays the user comments so far and also allows individual users to rate how much they like the drawing.

Templates and premium content. To showcase the creative possibilities of *12Pixels* we provide a special gallery of 'premium content' created by well known artists and designers. This serves as useful example imagery and a source of inspiration for users. Drawing templates such as speech bubbles, heart shapes, and face outlines are provided as an easy starting point for quick drawings. An additional 'Campaign' gallery is also provided for seasonal drawing challenges and official tie-ins.

Local save. To give drawings a life outside of the *12Pixels* application, drawings can be saved as standard GIF image files. This allows drawings to be utilized for any number purposes including mobile phone wallpaper, address book images, or sending them to friends via e-mail.

Creating emoji. One specific drawing usage for Japanese mobile phones is the creation of custom emoticons called *emoji*. Most Japanese phones have a pre-installed set of several hundred graphical *emoji* icons, but also allow for other custom icons to be embedded in rich-text email messages. We implemented a special *emoji* mode for *12Pixels* where users can create their own original *emoji* straight from their phone. *Emoji* are saved locally to the phones memory and can then be easily added to email for truly personalized communication. Figure 7 shows a selection of *emoji* created with *12Pixels* and an example of rich-text email usage.



Fig. 7. A selection of custom *emoji* (emoticons) created with *12Pixels* (left) and an example of their usage in rich-text email (right)

Personal manufacturing. As another way for *12Pixels* drawings to be shared between users a special service was created to manifest drawings into physical form. The 'crystal accessory' service turned user created drawings into custom fashion accessories as shown in Figure 8. By drawing on one of the templates provided, users would submit their design online to Ginza-based jeweler *Lights Style* and order original crystal jewelry such as dog tags, mobile phone stickers, and key holders.

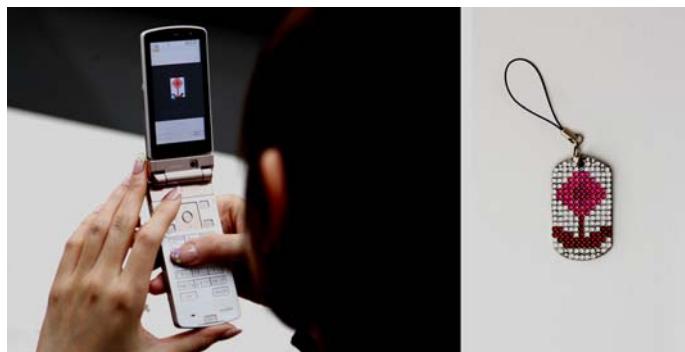


Fig. 8. Drawing with *12Pixels* to create an original accessory

5 Usage Analysis

User studies can be used to evaluate the efficiency of specific tasks or identify specific application weaknesses, but they do not necessarily indicate if a new idea will become successful or popular with the general public. Releasing a project to the wider population and observing its reception is the only true way to find such results. This is particularly true for Japan, which has developed a very unique mobile phone culture. Will people be interested in *12Pixels*? Will they be able to draw interesting images? What kind of images will they draw? How will they use the drawings? Without releasing *12Pixels* we believed it was all but impossible to convincingly answer these questions.

In this section we present analysis of the data collected following the public release of *12Pixels* in Japan. As *12Pixels* was released as an official service, we were fortunate to benefit from advertising and promotion of the application and quickly gathered a relatively large number of users. However this approach had certain disadvantages such as having to adhere to the Sony corporate privacy policy that did not allow us to access personal data or to directly contact end users for research interviews. Although we did collect a significant amount of data, we were only able to analyze it indirectly. This analysis is interesting and important for several reasons. Firstly, we discovered certain usage patterns for *12Pixels* that are applicable to a wide range of content creation applications. Secondly, our analysis suggests that creativity based applications and services are a promising direction for mobile phone development. Finally, we hope our analysis can inspire the development of new services, applications, and ideas for mobile creativity in general.

5.1 Data Analyzed

We made a comprehensive analysis of data from the first month of service after the initial release, from March 21 - April 21, 2009. The results of this analysis allowed us to re-think aspects of the *12Pixels* design and evolve new features for the subsequent release. Here we introduce this analysis and also make comparisons with data from the first month of the second release, June 30 - August 30, 2009.

5.2 Can People Draw with *12Pixels*?

Figure 9 shows a small sample of drawings created by *12Pixels* users since the initial release. It can clearly be seen that the variety of drawings is surprisingly extensive and complex; the quality and creativity of these drawings is nothing short of amazing.

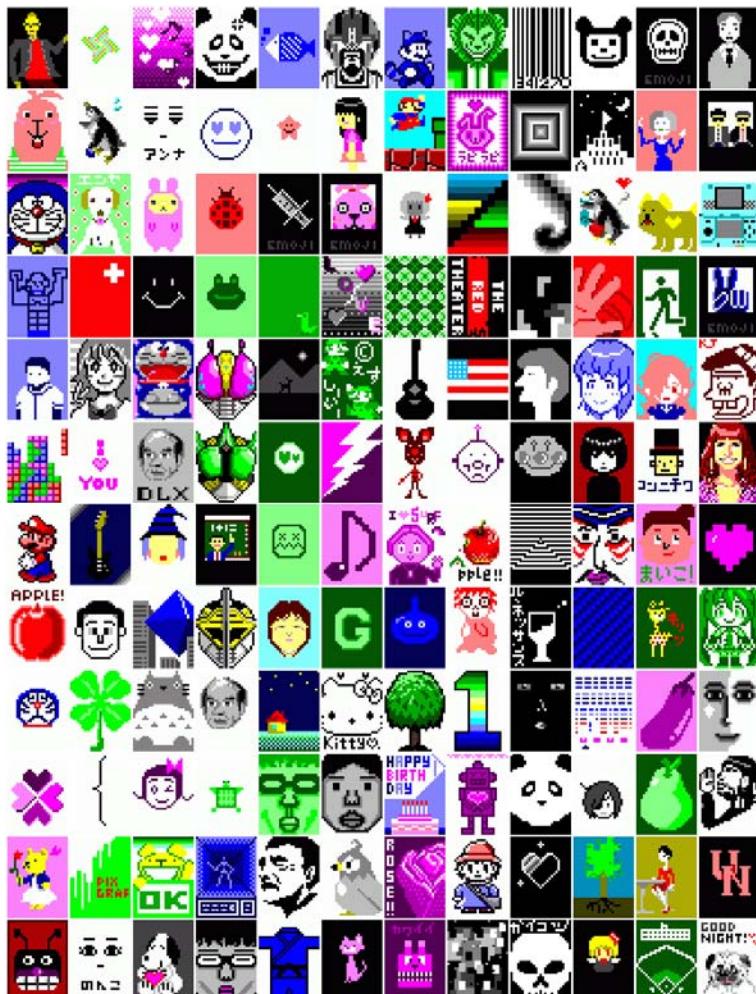


Fig. 9. A small sample of drawings created by *12Pixels* users in Japan

It is also apparent that technical drawing skills are not necessarily required to express oneself in interesting and unique ways. Indeed many of the simplest images are among the most expressive and enjoyable. Our experience with *12Pixels* demonstrates that like many user generated content services, social drawing can be a truly democratic medium that allows anyone to be creative.

5.3 Is 12Pixels Creatively Limiting?

In the process of designing and implementing *12Pixels* an often-expressed concern was the possibility of encountering a significant degree of repetition in user drawings. In other words, does the low resolution limit creativity?

The 27 x 36 pixel canvas offered by *12Pixels* may at first seem limiting, however the actual amount of unique imagery that can be created is very much vast. The vastness of the design space can be understood with a simple thought experiment. According to McCormack [17] if you were to spend every second of your entire life looking at a different variations of a black and white 6 x 6 pixel image, by the time of your death you would have seen less than 1% of all possible images. This suggests that the possibility space *12Pixels* provides is theoretically more than adequate for its intended purpose.

By analyzing the drawings we found the overwhelming majority were unique and special. There were however, some interesting exceptions. The first theme we noted was repetition when users drew *pervasive cultural symbols*, such as the heart mark or well-known characters. The second theme involved repetition when *remixing* and *repurposing* drawings from other users, this is discussed in more detail below.

These cases do not contradict our basic conclusion that low resolution and ‘enforced’ simplicity did not limit user creativity. From observing the thousands and thousands of drawings created since the initial release, there has been no indication that the canvas size became a limiting factor on user creativity. On the contrary, and as has been suggested by Boden [18], we found that the introduction of constraints on the creative process is what makes creativity possible by mapping out a territory of structural possibilities to be explored. This also creates a low entry barrier where users are less intimidated by the complexity of drawing and worry less about making mistakes.

5.4 Remixing and Repurposing

When we analyzed drawings uploaded to *12Pixels* over the first month, we observed that about 59% were original drawings, and the remaining 41% were repurposed original content.



Fig. 10. Examples of *remixed* (left), *recolored* (middle), and *progressive* drawings (right)

Within the 41% of repurposed content we found three distinct drawing sub-categories, example drawings are shown in Figure 10.

Remixed drawings (22%) are original drawings that were modified to produce a different variation by adding or removing various graphical elements.

Recolored drawings (4%) are drawings where only the color of the original drawing was changed without changes to the drawing itself.

Progressive drawings (12%) are uploads of a single drawing at various stages of development.

Other drawings (62%) are identical drawings uploaded multiple times by the same user or tests of the upload functionality such as blank images.

Analysis of drawing repetition demonstrated that 38% of drawings were re-used for purposeful reasons, as users actively modify drawings and contribute them back again. For the second release we were able to track how many times each drawing was downloaded to the *12Pixels* canvas specifically for repurposing (rather than just browsing). Of the 2276 drawings created in total, they were downloaded for repurposing 3790 times. This suggests that remixing and repurposing other drawings plays a major part in the creative cycle of *12Pixels*.

These observations lead us to rethink the distribution of the ‘premium content’ drawings we provided as examples. While this content was viewed frequently, it was not designed with remixing in mind and therefore was seldom appropriated. For the second release we shifted our focus to providing templates, rather than completed drawings, to accommodate easy remixing and repurposing.

5.5 What People Drew

What would people draw with *12Pixels*? We felt this question was important because it directly impacts upon the type of tools users require. To answer this question we surveyed drawings from the first month of the *12Pixels* release and identified the following most common image categories: *People* (11%), *Animals* (22%), *Pop Culture* (7%) i.e. artifacts commonly found in everyday culture, such as game characters, *Original Drawings* (7%), i.e. original characters; *Messages* (23%), i.e. drawings where text is the key element, *Objects* (7%), i.e. various objects such as cars, planes, etc, *Symbols* (20%), i.e. abstract or highly stylized imagery such as heart marks and *Others* (7%), i.e. drawings that do not fall into any of the above categories. Examples drawings from each category are shown in Figure 11.

While the above categorization is certainly an approximation, it provides an interesting insight into what and why people would draw on the mobile phone. It was quite unexpected that *Messages* represented the largest category with 23% of drawings. In



Fig. 11. Drawing categories: A: People, B: Animals, C: Pop Culture, D: Original Drawings, E: Messages, F: Objects, G: Symbols/Abstract

retrospect, it is a quite understandable – drawing messages creates highly personalized content that is both expressive and quick to create. This is particularly true for the Japanese language as it is very graphical by nature and lends itself to pictorial representation. *Animal* (22%) and *Symbols/Abstract* (20%) drawings follow closely, with the *Symbols* category prominent as it includes the heart symbol, which by itself constituted nearly 30% of drawings within the category.

Our analysis of the drawings suggests that one of the most important reasons for people to draw on the mobile phone was to *support communication*, rather than simply express oneself creatively. This trend is even more evident when we estimated how drawings were used once created.

5.6 How Drawings Are Used

While we cannot track how people use drawings locally on their phone, e.g. as custom wallpaper, we can track the use of *12Pixels* services to create *emoji* and physical accessories. From the first release 45% of the images uploaded to the communal gallery were *emoji* icons and roughly 4% of images were drawn to create physical accessories. For the second release, *emoji* icons accounted for 39% of drawings uploaded during the same one month period.

This observation indicates that a large percentage of *12Pixels* drawings were used as a means to support self-expression when communicating via email. Hence, *12Pixels* had significant use as a tool for *supporting communication*, rather than creative expression alone. We were somewhat surprised that creating physical accessories did not prove very popular with users. The expense of ordering the accessories is one possible reason, and it is also possible that *12Pixels* is viewed as a casual drawing tool rather than for ‘serious’ accessory design.

5.7 Usage Patterns

In highly urban areas such as Tokyo, long periods are spent with mobile phones while commuting to and from work, a 2005 survey calculated the average one-way commute time to be an astonishing 67 minutes for workers and 72 minutes for students [19]. We envisioned *12Pixels* as a tool to be used during these ‘downtimes’ on the train to and from work, while waiting for friends, or in any other in-between times. From the data collected from the first release we analyzed drawing upload times to gain a better picture of usage patterns. Figure 12 shows the distribution of uploads by hour over the course of the day. We can see that the peak upload time occurs when people are commuting home after leaving work around 6pm. The second peak comes after dinner at around 9pm and another lesser peak around lunchtime. While not a definitive answer to the question of when and where people are using *12Pixels*, this data provides a useful insight. It in part substantiates our assumption that *12Pixels* would be used during small pockets of downtime, however we were surprised to find that there was sparse activity during the morning commute and continued activity after returning from work.

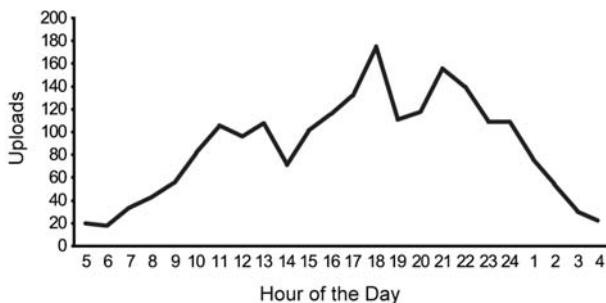


Fig. 12. The distribution of drawing uploads over the course of the day

5.8 Evolution of *12Pixels*

The second release of *12Pixels* has included a number of improvements driven by the analysis that we presented above. A refined color menu interface was added, the process of creating *emoji* was significantly improved, content available online was rethought to cater for remixing, and we also significantly strengthened the social capabilities by adding a commenting and rating system.

The result of these improvements was a significant increase of drawings uploaded to the server. In the first month of use, the number of drawings uploaded increased by approximately 300% from 780 to 2276 in the second release. The number of repeat users, i.e. those who uploaded 2 or more drawings, increased from 42.8% to 61%. This data supports the validity of our observations and shows a growing real-world interest in *12Pixels*.

6 Discussion

With the current proliferation of social web applications there appears to be genuine and growing interest in all forms of computer based creativity and content creation. Through the *12Pixels* project we have explored the approach of ‘social drawing’ on the mobile phone and uncovered new ways for people to express themselves. This section presents some of the conclusions drawn from our experiences.

6.1 Mobile Phones for Creativity?

One of the main questions our research sought to answer was: *Are people interested in using their mobile phone creatively and in particular for drawing?* We believe that our observations point to a positive answer to this question. Indeed, the drawings and creative output that we observed from thousands of people surprised and amazed us. At the time of this writing nearly 25,000 drawings have been contributed to the *12Pixels* gallery in under a year. This suggests that using *12Pixels* is an engaging and appealing process. We believe that tools for creative expression on the mobile phone provide a unique and special experience that is not found on traditional desktop computing platforms.

6.2 Simplicity vs. Expression

User discussions and observations suggest that the *simplicity* of the drawing interface was one of the key reasons behind *12Pixels* appeal. Throughout the process of developing *12Pixels* we were constantly forced to consider the trade off between ease of use and expressive ability. It was critical to find the appropriate balance; users frustrated by the difficulty of interaction will soon give up, as will users bored by the limited possibilities of expression. In the words of Levin we strived to create an instrument that like the pen and the piano is ‘instantly knowable, indefinitely masterable’ [20].

Have we been able to achieve this balance with the *12Pixels* interface? Based on the amazing quality of drawings created, we believe we have come close. One memorable comment received from an interaction designer working in the mobile phone industry described *12Pixels* as being: ‘Simple to use, difficult to draw’. This may seem contradictory but it very much echoes Levin’s maxim above. As deceptively simple as the pen and the piano are, without mastery there can be no masterpiece.

We believe that striking the balance between expressiveness and ease of use is a crucial challenge for developing any interface for creative expression. Achieving this balance requires careful consideration of the creative process, and its immediate context. We found that the technical constraints of the mobile phone, while challenging, can be utilized to define a very specific area of creative activity. In our case using the specific style of pixel art may at first seem like a major limitation, however it in effect acted as a mechanism to free up the creative process by posing well considered constraints and encouraged participation.

6.3 Creative Communication

We found that one of the most important reasons why people would use *12Pixels* was to enhance communication. This was evident in the number of drawings created with message based content and in the surprisingly high percentage of *emoji* icons created for personalized email communication. We believe the role of communication should be carefully considered for any usage of the mobile phone as a creative platform – adding communication capabilities is extremely important. Exploring the link between creativity, communication, and mobility can be a fruitful and exciting area of research for the pervasive computing community.

7 Future Directions

We believe that developing applications to support everyday creativity is an exciting and promising area of future research for pervasive computing applications. The *12Pixels* project explored only a limited range of the possibilities for mobile creativity. Areas we have yet to explore include developing an enhanced drawing interface for animation, with functionality to create and share quick animated sequences. Drawing interfaces can also explore the use of the mobile phone camera to capture color, provide images for annotation, and develop simple digital image processing techniques.

As the functionality of the mobile phone extends to include more and more web based services, there are numerous opportunities for enhancing them with *12Pixels*-style drawing functionality. For example drawings can be used as custom ‘visual tags’

to mark out distinct locations on an online map or even be used as input in a visual search engine. Drawing is a form of expression that has to-date been largely ignored by mobile phone research and there remain a surprisingly wide range of innovative applications yet to be developed. We hope this work will encourage researchers and practitioners to explore this exciting area.

Acknowledgments. For their help in bringing *12Pixels* to market we thank Testu Natsume from Sony CSL, Seimi Tezuka, Tomohiro Naito, Masaki Maeda, and Hiroki Yamaguchi from Sony Marketing, Yuji Mizuto, and Yosuke Fukuawa from Imaginative Inc.

References

1. Rideout, V., Roberts, D.F., Foehr, U.G.: Generation M: Media in the Lives of 8–18 Year-olds. A Kaiser Family Foundation Study (2005)
2. VeronisSuhlerStevenson: VSS Communications Industry Forecast 2009–2013 (2009)
3. Poupyrev, I., Willis, K.: TwelvePixels: Drawing & Creativity on a Mobile Phone. In: CHI 2008. ACM SIGCHI (2008)
4. Gartner: Market Share for Mobile Devices (2009)
5. Goodyear, D.: I Love Novels. The New Yorker (2008)
6. Kurosu, T.: Development of Mobile Internet and the Change of Young People's Information Behavior (2008)
7. Yoshitaki, S., Ohta, M., Kawaguchi, A., Ishihara, S., Mizuno, T.: CONTE: A Drawing Tool for Mobile Phones. IPSJ SIG Notes. MBL 24, 211–218 (2002)
8. Yoshitaki, S., Ohta, M., Sakane, Y., Ishihara, S., Mizuno, T.: CONTE: A Drawing Tool for Free-form Curves on a Mobile Phone. Transactions of Information Processing Society of Japan 44, 285–296 (2003)
9. Jingtao, W., John, C.: TinyMotion: camera phone based interaction methods. In: CHI (2006)
10. Jenkins, H.: Confronting the Challenges of Participatory Culture: Media Education for the 21st Century (2006)
11. Papert, S.: Mindstorms: children, computers, and powerful ideas. Basic Books, New York (1980)
12. Resnick, M., Silverman, B.: Some Reflections on Designing Construction Kits for Kids. In: Interaction Design and Children Conference (2005)
13. Beverly, L.H., Kenneth, P.F., Anuj, G., Carlos, M., Roy, W.: Squeeze me, hold me, tilt me! An exploration of manipulative user interfaces. In: CHI (1998)
14. Rekimoto, J., Ishizawa, T., Schwesig, C., Oba, H.: PreSense: interaction techniques for finger sensing input devices. In: UIST (2003)
15. Robbins, D.C., Cutrell, E., Sarin, R., Horvitz, E.: ZoneZoom: Map Navigation for Smartphones with Recursive View Segmentation. In: AVI. ACM, Gallipoli (2004)
16. Dearman, D., Inkpen, K.M., Truong, K.N.: Target Selection on Mobile Devices using Display Segmentation. In: Mobile HCI. ACM, Singapore (2007)
17. McCormack, J.: Facing the Future: Evolutionary Possibilities for Human-Machine Creativity. In: Romero, J., Machado, P. (eds.) The art of artificial evolution: a handbook on evolutionary art and music, pp. 417–451. Springer, Berlin (2008)
18. Boden, M.A.: The Creative Mind: Myths & Mechanisms. Weidenfeld and Nicholson, London (1990)
19. Foreign Press Center Japan: Facts and Figures of Japan - Housing (2005)
20. Levin, G.: Painterly Interfaces for Audiovisual Performance Massachusetts Institute of Technology (2000)

No-Look Notes: Accessible Eyes-Free Multi-touch Text Entry

Matthew N. Bonner, Jeremy T. Brudvik, Gregory D. Abowd,
and W. Keith Edwards

GVU Center & School of Interactive Computing, Georgia Institute of Technology
85 5th Street NW Atlanta, GA 30308 USA
`{matt.bonner,jbrudvik,abowd,keith}@gatech.edu`

Abstract. Mobile devices with multi-touch capabilities are becoming increasingly common, largely due to the success of the Apple iPhone and iPod Touch. While there have been some advances in touchscreen accessibility for blind people, touchscreens remain inaccessible in many ways. Recent research has demonstrated that there is great potential in leveraging multi-touch capabilities to increase the accessibility of touchscreen applications for blind people. We have created No-Look Notes, an eyes-free text entry system that uses multi-touch input and audio output. No-Look Notes was implemented on Apple's iPhone platform. We have performed a within-subjects ($n = 10$) user study of both No-Look Notes and the text entry component of Apple's VoiceOver, the recently released official accessibility component on the iPhone. No-Look Notes significantly outperformed VoiceOver in terms of speed, accuracy and user preference.

Keywords: accessibility; mobile device; multi-touch; touchscreen; text entry; eyes-free.

1 Introduction

The development of touchscreens sensitive to multi-finger input has sparked a renaissance of interest in the technology's popularity. Devices like the Apple iPhone are part of a rush to take advantage of this wave of interest. Touchscreens also bring undeniable utility, reducing the need for peripherals like keyboards and lending themselves to collaborative use. Effective and flashy, touchable interfaces have appeared at the cash register of the supermarket, at the check-in line of the airport, and in the pockets of the masses.

Unfortunately, as discussed by McGookin *et al.* [1], the creation of accessible modifications and enhancements for touch-based devices is lagging behind the breakneck pace of mainstream development. Touchscreens pose an especially daunting challenge for blind or visually impaired users. In place of familiar devices like the keyboard, screen computing generally offers a uniform, featureless surface. The trend of emulating existing GUI's while using the finger as a mouse can translate existing problems into an even more difficult context. On top of

this, mobile systems often feature unique or restructured interface layouts. The end result is a confounding environment for accessibility programs.

Eyes-free use by sighted users is also difficult. Touchscreens demand the user's visual attention, making it extremely difficult to carry out other tasks simultaneously. In some cases, social norms require discreet or covert use of a touchscreen. For example, a meeting participant could respond to an urgent text message or jot down a note without appearing inattentive. There are even situations where focusing on a touch screen is physically hazardous, such as sending a text message while driving.

On a system like the iPhone, users enter text for a myriad of tasks including text messaging, internet browsing and the use of third-party applications. Text entry proceeds with a virtual keyboard, creating a plethora of miniature, adjacent targets that are completely indistinguishable from one another without visual examination. Competitors like the Blackberry Storm by Research In Motion have followed suit with their own virtual keyboards. Text entry is thus a primary means of interacting with devices like the iPhone and Blackberry Storm, but it is extremely difficult to do without visual input.

Following this conviction, we developed No-Look Notes, an eyes-free gesture-based text entry system for multi-touch devices. No-Look Notes offers two-step access to the 26 characters with a small number of simple gestures that remove the precise targeting required by soft keyboards. Some basic text editing actions, such as backspace, are included as well. The combination of gestures and multi-touch capabilities results in a system that is both exploratory and expeditious.

We begin this paper with a discussion of related work and a formative pilot study. We then present the design principles that follow from this synthesis, and describe the design of No-Look Notes. Next we report the results of a user study with visually impaired subjects, testing both No-Look Notes and the text entry component of Apple's new eyes-free accessibility tool, VoiceOver. We conclude with an analysis of our results and discuss future work.

2 Related Work

Text entry on mobile devices is a well-studied research area. Mackenzie and Soukoreff give a thorough overview of text entry systems on mobile devices using pen-based and keyboard-based input [2]. Our system, No-Look Notes, focuses on a third paradigm of input: multi-touch text input. Eyes-free multi-touch text entry is an exceptionally young research topic, so our related work draws on both single and multi-touch text entry work, as well as the haptic augmentation of touchscreens.

2.1 Eyes-Free Single-Touch Text Entry on Touchscreens

Sánchez and Aguayo developed “Mobile Messenger for the Blind” [3], a messaging system for mobile devices that divided the screen into a 9-button virtual keyboard with multiple characters on each virtual key, much like a mobile phone

keypad. To specify a character, a key must be tapped multiple times. The system used text-to-speech (TTS) for output.

While it is clearly difficult for blind people to perform positional input, such as using a computer mouse or finding targets on a touchscreen, Mobile Messenger for the Blind made this task easier by keeping the number of targets low, placing them at easy-to-reference locations based on the physical characteristics of the device (*e.g.*, the edges and the corners of the screen), and keeping the targets static.

Increasing the number of targets will not only make a target-based system more difficult to explore, but also more difficult to enter text on. Our system, No-Look Notes, completely avoids the accuracy issues of requiring users to precisely tap several times on a target by using a multi-touch input system. By using a circular layout, No-Look Notes may also more easily add new input ‘targets’ while maintaining a simple exploration strategy.

Numerous systems rely on gestures for text entry, sometimes in combination with targets. Tinwala and MacKenzie developed a system based on Graffiti strokes for eyes-free text entry [4]. While the system is targeted at non-visual use, Graffiti strokes require the user to trace complicated forms (*e.g.*, english letters) onto the screen. Ken Perlin’s Quikwrite [5], though not targeted at visually impaired users, attempts to speed Graffiti entry by replacing letter shapes with regions. Users enter text by dragging to a group of letters present on a particular screen region, then to a secondary subregion to enter a specific letter. As described by Plimmer *et al.* [6], writing and learning to write is an extremely challenging task for visually impaired users. It is unlikely that a Graffiti-based system, or a system involving multiple precise and sequential gestures, will be usable for many visually impaired users.

Yfantidis and Evreinov created a system that used extremely simple unidirectional single-finger directional gestures that mapped onto pie-menus of characters [7]. A gesture in any of the 8 compass directions (North, Northeast, East...) corresponded to a unique character. 24 characters were mapped onto 3 separate “layers,” or pie-menus, which are traversed based on the delay of the finger between touching the screen and commencing a gesture. This system also strictly used TTS for output.

The gesture-based approach of this system does not require the user to hit any targets. The time-delay to switch between screens could be improved upon by introducing multi-touch interaction techniques, but this was a hardware limitation and not a system implementation issue.

At Google, T. V. Raman has created a system that incorporates elements of both gestures and soft buttons [8]. As in Yfantidis and Evreinov’s system, characters are arranged into ‘rings’ which are selected by a gesture towards the desired character. Two representative letters from each ring are arranged on an additional default ring. Users enter letters by first selecting one of these representative characters, which transitions the screen to a secondary ring. On this secondary ring, users select a character, and begin the process anew. This design speeds text input by removing the need to wait to move between menu screens.

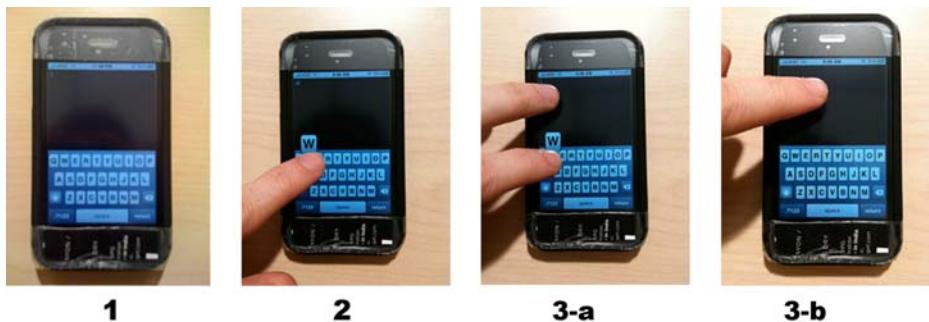


Fig. 1. Entering a character using VoiceOver: **1** Basic ‘soft’ QWERTY keyboard. **2** Rest finger on desired character. **3-a** Tap screen anywhere with second finger (‘split-tap’) or alternatively. **3-b** Lift up first finger and then double-tap the screen anywhere.

Both Yfantidis and Raman attempt to avoid the issue of accurately tapping by using gestures and have explicitly chosen simple, straight-line gestures. Despite this, accurately gesturing without visual feedback is challenging. Users must also first ensure their gesturing finger is oriented properly on the phone. No-Look Notes addresses the accuracy issues of targets by adding multi-touch instead of switching to gestures, thereby avoiding these potential pitfalls.

The largest problem with these two systems however is their lack of an effective exploration method. Users must continually swipe in different directions until the desired letter is found, deleting the character every time it is incorrect. After entering a letter, users must distinguish that series of swipes and delays from all the previous erroneous but highly similar input gestures. Introducing a completely novel and non-trivial organization of letters, as in Raman’s system, could further the confusion. Because it uses a multi-touch technique for text entry, No-Look Notes is able to reserve single-fingered interaction for exploration. Users may drag their finger around the screen seeking a letter without entering characters along the way.

2.2 Multi-touch Interaction

During No-Look Note’s development, Apple announced “VoiceOver,” a system for eyes-free use of the iPhone that relies on touch-input and TTS output [9]. VoiceOver’s general scheme uses a ‘focus’ box which is moved between UI elements either by touching an element or flicking a finger in the direction the user wants focus to move. Double-tapping the screen activates the item currently inside the focus box. “Split-tapping,” in which the user first rests a finger on the desired item and then touches the screen with a second finger, may also be used to select an item. An early example of this general strategy of providing an ‘audio overlay’ for an unaltered GUI by moving focus between UI elements was demonstrated by Mynatt and Edwards [10].

Text entry in VoiceOver is subsumed by this basic focus-box system. Users interact with a soft keyboard (and any other application) by using this

system: touching near the key they wish to use, ‘swiping’ to move the focus box until it is correctly targeted if necessary, and split-tapping or double tapping to enter the key (see Figure 11). VoiceOver has additional functionality for examining the entered text. Users may swipe a finger up or down on the screen to move a cursor/insertion point through a line of text, reading each character the cursor moves over. In the rest of this paper, when referring to “VoiceOver,” we refer only to this text entry component.

Because it relies on a QWERTY keyboard layout, VoiceOver presents the user with a large number of targets. This can make it difficult to locate a letter, even though the double-tapping/split-tapping system helps avoid target-tapping accuracy issues. Additionally, the screen will be a non-sensical jumble of letters to users unfamiliar with computer keyboards.

No-Look Notes also uses split-tapping to allow the use of targets while avoiding their potential for accuracy issues. We further ease exploration by minimizing the number of targets on the screen at one time, using a simple alphabetical character-grouping scheme based on phone keypads.

As Apple’s official accessibility system included by default on the latest iPhone versions, VoiceOver will be distributed as the *de facto* accessible text entry system for a very large number of users. This makes VoiceOver both the first publicly released and first widely distributed multi-finger text entry system for touch screens. No-Look Notes was extensively and directly compared against VoiceOver in our evaluation, in which we evaluate both systems.

VoiceOver’s “split tap” is identical to the “second-finger tap” developed in Slide Rule by Kane *et al.* [11]. Slide Rule is a set of multi-finger interaction techniques for list-based applications, such as a phonebook and a music player. A complementary pair of techniques developed were the *one-finger scan* and *second-finger tap*, where sliding one finger on the screen is used to browse lists and a second finger tap, while still touching the screen with the first finger, selected an item in the list.

This approach both promotes risk-free exploration of items and allows blind users to select items without hitting a target on the screen. It is clear from this system that relatively simple multi-finger interaction techniques can greatly improve the accessibility of touchscreens. Our system also employs the split-tap from Slide Rule and applies it to text entry.

2.3 Haptics and Touchscreens

Haptic feedback for touchscreens tends to appear as an augmentation of button-based interfaces, as in work by Brewster [12], Leung [13] and Kaaresoja [14]. Brewster in particular showed simple haptic augmentation was beneficial for entering text on a small virtual keyboard. Recent work by Yatani and Truong explores other uses for haptics, using vibration to assist a gesture-based system [15].

Our system eschews haptic feedback because we wanted to develop a system that could be used on any touchscreen without modification. Touchscreens not integrated in mobile devices (*e.g.*, mobile phones) are unlikely to have vibration

motors included. Systems like Yatani’s require extensive modifications even to a phone that offers vibration.

3 Design

3.1 Pilot Study

The design of No-Look Notes was informed by both the aforementioned work and by a pilot study of a prototype system. This pilot system was our first attempt at using multi-touch interaction to improve a gesture-based system. In our prototype, letters and letter groups (grouped by frequency of predicted use) were arranged in pie menus around the screen and as in the systems by Yfantidis [7] and Raman [8].

Rather than cycling through the pie-menus by selecting a representative character or adding a time delay, we tested a multi-touch interaction technique. By resting a finger on the screen, the user switched to a second pie menu. Users thus entered a letter or letter group by swiping in a letter’s direction, or resting a finger on the screen and then swiping.

We tested our prototype with five visually impaired participants, each of whom entered text for about 1.5 hours. This testing exposed three interrelated qualities that are absolutely key for a successful eyes-free text entry system: (1) Robust Entry Technique, (2) Familiar Layout, (3) Painless Exploration.

Robust Entry Technique. Eyes-free systems must provide text entry techniques that are both exceptionally simple and exceptionally error-tolerant. Without visual confirmation, it is difficult for users to make precise gestures or hit precise targets. Mundane interaction techniques rapidly increase in complexity when used in an eyes-free context. Small distinctions, such as the difference between gesturing ‘Left’ and ‘Up-and-Left’ or the difference between a full circle and a three-quarters circle, are especially difficult.

Familiar Layout. A layout that is easy to conceptualize and related to familiar interfaces or groupings is critical for an eyes-free system. Because touch-screens and multi-touch gestures are already foreign to most visually impaired users, an interface needs to include a recognizable layout it can to reduce the cognitive load on its users. Alphabetization, for example, will be far more successful than a layout based on theoretical character frequencies.

Painless Exploration. Users of an eyes-free system must be able to painlessly explore the system’s layout, not just correct their mistakes. Exploring by repeatedly entering and undoing actions is not acceptable, there must be some first-class exploration technique. “Entering and undoing” adds to the user’s cognitive load by requiring them to remember which entry was correct among many similar entries. Painless exploration is aided by a familiar layout and robust entry technique, but does not necessarily follow from these.

3.2 No-Look Notes

No-Look Notes arranges characters around the screen in an 8-segment pie menu reminiscent of the systems proposed by Yfantidis and Raman. However, each section of the menu contains multiple characters, such that all 26 letters of the English alphabet appear (see Figure 2). The 8 character groups (*e.g.*, ‘ABC’, ‘PQRS’) correspond to the international standard mapping a phone keypad to letters (Familiar Layout) [16].

The segments are soft-buttons which must be touched. When the user touches a segment, either by dragging their finger to a new segment or touching the screen, the characters in that segment are announced audibly. The user may drag and tap their finger around the screen (for example, tracing the screen’s edges) without accidentally entering characters (Painless Exploration).

Resting one finger on a segment and tapping a second finger (*i.e.*, split-tapping or second-finger tapping) selects that segment, bringing the user to a secondary screen with that segment’s characters from that selection arrayed alphabetically from the top to the bottom of the screen. Users select the desired character the same way they selected a character group. The user drags a finger until they hear the desired character announced, then drop a second finger to the screen to select (Robust Entry Technique, see Figure 3).

This simple arrangement makes it fast, easy, and risk-free to search for a character: simply drag a finger around the screen. Users may trace the edges of the screen and eventually reach any character, since every ‘pie slice’ reaches the edge of the screen. This also allows No-Look Notes to leverage some of the “pure” benefits of edges, as defined by Wobbrock *et al.* [17], such as higher accuracy.

In addition to character entry, No-Look Notes offers gestures for “space” and “backspace/undo.” Backspace is a quick swipe with one finger to the left, space is a quick swipe to the right. The backspace gesture is also used to cancel a selected character group without entering a character. Reading (TTS) and spelling currently entered text is triggered by a swipe down. These swipes require a minimum distance and speed in order to register, preventing a user exploring the screen by tapping or dragging their finger from accidentally activating them.

4 Evaluation

We performed a within-subjects evaluation of No-Look Notes and VoiceOver. Users spent 15 minutes learning a system in an interactive tutorial with an experimenter, then spent 1 hour using the system to enter words, and finally answered a brief questionnaire. The second system was then tested in the same way. Users were split into one group evaluating No-Look Notes first and another evaluating VoiceOver first to counterbalance order effects.

MacKenzie and Soukoreff deal extensively with evaluation in their treatise on mobile text entry [2]. The value of quantitative *and* qualitative results from users is emphasized. Measurement using words-per-minute, as well as examining the relationship between speed, accuracy and errors are also covered.



Fig. 2. No-Look Notes. The screen on the left is the main screen. The screen on the right is after the ‘ABC’ target is selected from the main screen. Visual representation was added to this figure for illustration.

4.1 Participants

We recruited 10 participants, 7 men and 3 women. Participants were recruited from the Atlanta Center for the Visually Impaired. The average age of participants was 40.8 ($sd=10.85$). All users were visually impaired, where visually impaired is defined as “requires assistive technology (screen readers, magnification) for computer use.” No participant was able to visually distinguish specific keys or letters on the phone screen for either system. Additionally, no participant could view the text they had entered.

4.2 Device

Participants entered text on an Apple iPhone, which has a 3.5 inch capacitive touchscreen. Non-touch sensitive points on the top and bottom of the phone (near the ear and mouthpiece) were taped over to give the screen tactile boundaries.

No-Look Notes was implemented as an iPhone OS 3.0 application. TTS for characters and actions (like backspace) was pre-synthesized using Mac OS X’s built-in TTS engine and loaded onto the phone as audio files. VoiceOver’s release introduced dynamic TTS on the iPhone, but developers do not have access to this functionality. No-Look Notes thus provided dynamic TTS using Flite [18].

VoiceOver’s iPhone OS 3.0 version was tested using a custom application to isolate the text entry portion of the system. No modifications to the actual text entry or TTS were made. However, target words were synthesized by Flite.

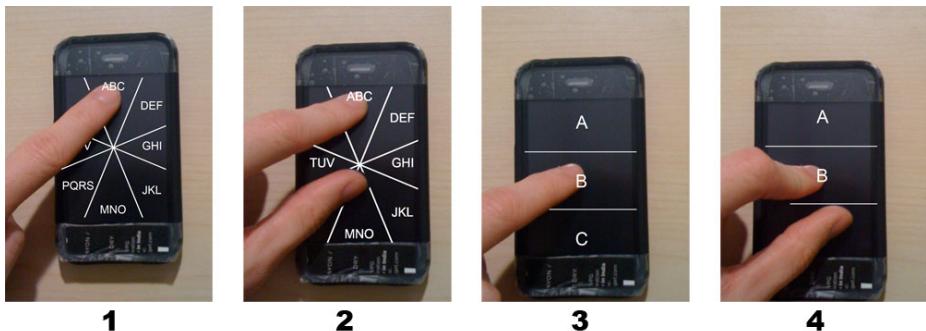


Fig. 3. Entering a character using No-Look Notes: **1** Rest finger on desired group **2** Tap screen anywhere with second finger ('split-tap') **3** Rest finger on desired character **4** Split-tap. Visual representation was added to this figure for illustration.

Participants tested VoiceOver in 'portrait' mode, with the keyboard taking up the bottom half of the screen.

4.3 Procedure

After their introduction to the iPhone's physical characteristics, participants ran the following procedure for each system. To avoid order effects, we counterbalanced participants such that half started with No-Look Notes and half started with VoiceOver.

First, participants spent 15 minutes learning the system. Experimenters taught the participants how to explore, enter and delete characters, as well as how to read characters on the screen. Participants were required to demonstrate their grasp of each new action, and encouraged to ask questions. Prior to beginning testing, participants demonstrated their competency by entering a test word.

Following this practice session, participants spent 1 hour entering single words from a published phrase set for text entry, also by MacKenzie [19]. This was selected over requiring participants to memorize phrases to reduce the cognitive load on the user, as also discussed by MacKenzie [2]. Phrases were randomly chosen from the phrase bank to create a phrase set; this same phrase set was used by all participants for each system.

Timing begins after the user enters a character. Each input (including errors) is duly noted by the system. An incorrect character entry caused an error sound to play, and the user was required to backspace/delete the offending characters. Timing continues until the moment the user's entered text completely matches the target word. At this point, the timer pauses and the next target is queued, waiting for the user to touch the screen.

Participants were able to rest whenever they wished, but were encouraged to rest between words rather than mid-word. Participants were also reminded several times that this was a test of the system, not their speed or spelling skills.

After completing 60 minutes of text entry, participants responded to a brief questionnaire about their opinions of the system used in that condition. Participants then repeated the 15-minute practice/learning + 60 minutes of use + questionnaire cycle with the other system.

5 Results

Overall, we collected 20 total hours of usage data (1 hour per system * 2 systems per participant * 10 participants) counting rest time between words, which yielded a total of 3921 correct characters and 1135 correct words entered.

5.1 Physical Comfort

Some participants mentioned that their hands or fingers were tired. One participant noted that his hands were “tingly.” Although some users required minor breaks, all users were able to complete the entire study.

5.2 Text Entry Speed

Due to technical limitations of the iPhone SDK, we were unable to access gesture information for the VoiceOver condition. Therefore, we started timing of target words at the first input we could measure in the condition: after a character was entered. The same timing was used for the No-Look Notes condition. Since we do not time until after the first character is entered, we considered target length to be $n - 1$ when calculating the text entry speed measure. We used the WPM (words-per-minute) measure for text entry speed, calculated as (characters per second) * (60 seconds per min) / (5 characters per word). This timing technique and text entry speed measure are identical to those used by Wigdor and Balakrishnan [20]. While we chose to consider a “word” to be 5 characters in this analysis, in actuality, the average length of the 1135 target words entered was 3.45 characters.

Although we counterbalanced the order conditions across participants, it is still possible to encounter asymmetric transfer effects. To test this, we performed one t-test on each system’s text entry speed for each of the two groups formed by counterbalancing condition order (No-Look Notes first, VoiceOver first) for two t-tests total. The results of these tests were non-significant (No-Look Notes speed: $p = 0.50$, VoiceOver speed: $p = 0.88$) which suggests that there was no order effect.

The overall text entry speeds were 0.66 WPM for VoiceOver and 1.32 WPM for No-Look Notes, a 100% increase in favor of No-Look Notes. This difference was determined to be significant using a paired t-test ($p < 0.001$). Figure 4 shows text entry speed performance for each user for each system; all but one participant achieved a higher speed with No-Look Notes.

To examine how participants’ speed varied over the course of the session, we split up each hour into 6 10-minute blocks and calculated speed within each

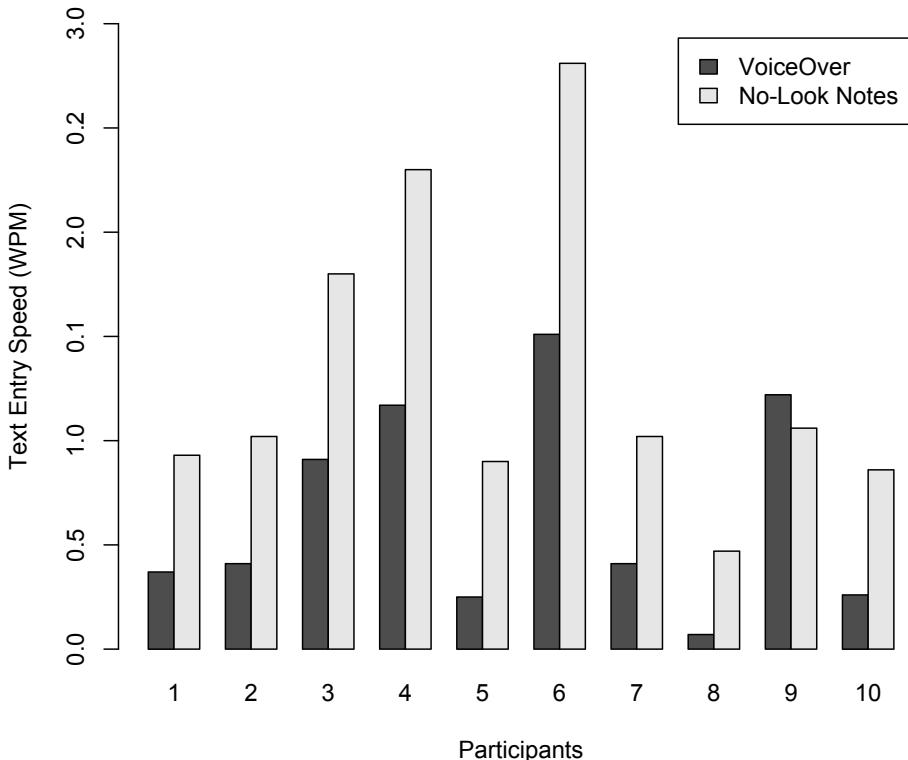


Fig. 4. Text entry speeds (WPM) for each participant for each system

block. The averages across participants for block-level speed are shown in Figure 5. Text entry speeds for the first block for VoiceOver and No-Look Notes were 0.85 WPM and 0.61 WPM, respectively. The maximum text entry speed for each system occurred in the fourth block: 0.76 WPM for VoiceOver, 1.67 WPM for No-Look Notes. While these text entry speed curves may fit a general learning curve with many sessions over time, the participants' speed declined towards the end of our sessions, possibly due to fatigue.

5.3 Text Entry Errors

We calculate the error rate as (incorrect characters entered) / (correct characters in target word), as is also done by Widgor and Balakrishnan [20].

As with text entry speed, we tested for the occurrence of order effects for error rate. Neither order was significant for either condition (No-Look Notes error rate: $p = 0.71$, VoiceOver error rate: $p = 0.62$), so the performed counterbalancing was also acceptable here.

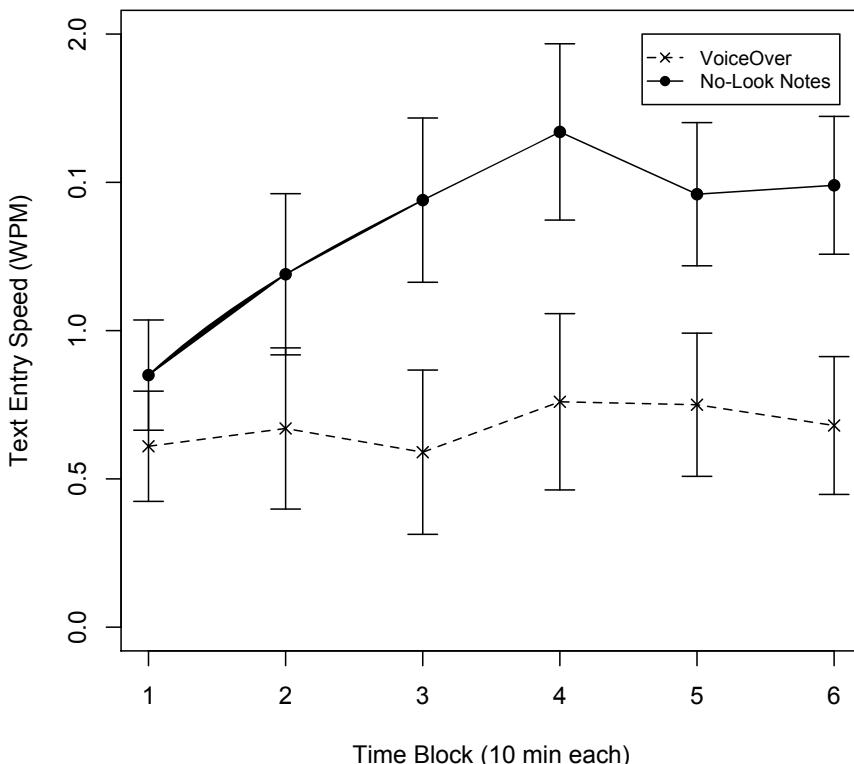


Fig. 5. Text entry speed (WPM) across participants for each 10-minute block (time interval)

The overall mean error rates across participants were 0.60 for VoiceOver and 0.11 for No-Look Notes, a 44% increase in errors for VoiceOver over No-Look Notes. Using a paired t-test, this difference was found to be significant ($p < .05$). Error rates for each participant for each system are shown in Figure 6.

5.4 Questionnaire Results

After each condition, participants responded to a brief questionnaire about their opinions of the system used in that condition. The questionnaire comprised 14 statements in which the participant would state their agreement on a 5-point Likert scale (1 = disagree strongly, 5 = agree strongly). All of the mean responses were higher for No-Look Notes than for VoiceOver. Using a paired Wilcoxon test, we determined that 8 of 14 differences were significant ($p < .05$), all in favor of No-Look Notes, while 6 of the 14 were not significant. The list of statements, mean responses, and significant differences are shown in Table II.

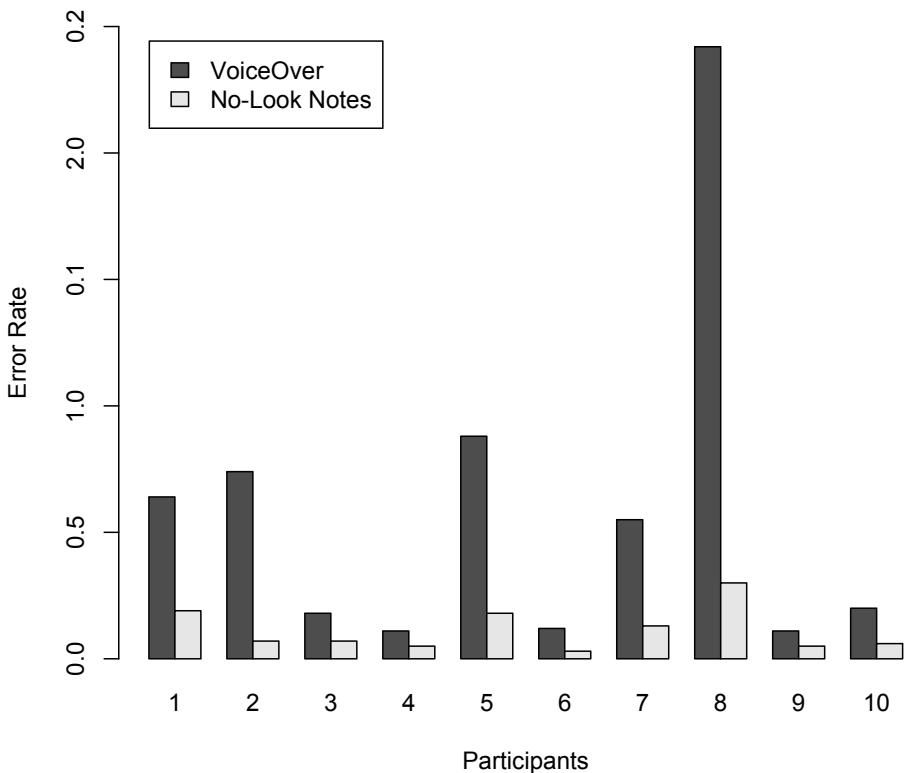


Fig. 6. Error rates for each participant for each systems

No-Look Notes was rated significantly higher than VoiceOver for the majority of statements, including key statements such as “easy to learn,” “fast to use” and “felt in control.” Our questionnaire results suggest that most users felt No-Look Notes was both easier to learn and faster overall. This supports our speed and error monitoring results, which shows that this was actually the case.

5.5 Qualitative Feedback

Seven participants responded negatively to the size and number of targets in VoiceOver, finding it difficult to locate keys. Participant 3 said *“If I wasn’t familiar with the QWERTY [keyboard layout], it would have been hell”*. Indeed, participant 8, our only participant with no QWERTY experience, achieved a rate of only 0.07 wpm with VoiceOver, saying mid-use: *“I want to cry right now.”*

Participants responded positively to the familiar aspects of both systems. Participants with knowledge of QWERTY keyboards (nine out of ten) said that thinking of the keyboard layout aided them in locating keys in VoiceOver. The fact that groups of characters in No-Look Notes are the same groups used on a phone keypad was also mentioned favorably by several participants. It was also

Table 1. Questionnaire results (mean, sd) for responses given on a Likert scale (1 = disagree strongly, 5 = agree strongly). The mean value for each question favored No-Look Notes; asterisks indicate where this difference was significant.

Statement	VoiceOver	No-Look Notes
Easy to use*	2.6 (1.43)	4.2 (0.63)
Fun to use	3.4 (1.71)	4.7 (0.67)
Fast to use*	3.0 (1.83)	4.4 (0.70)
Felt in control*	3.0 (1.63)	4.2 (0.79)
Easy to learn*	3.2 (1.48)	4.6 (0.70)
Intuitive	3.4 (1.43)	4.4 (0.97)
Familiar	2.7 (1.57)	3.4 (1.26)
Features clear to me*	3.7 (0.95)	4.5 (0.71)
Improve with practice*	3.7 (1.25)	4.8 (0.42)
Would use this system	3.9 (1.37)	4.7 (0.67)
Made entering text accessible*	3.7 (1.25)	4.8 (0.42)
Aware of text I was entering*	2.8 (1.40)	4.5 (0.85)
Audio feedback clear	3.2 (1.55)	3.3 (1.70)
Easy to undo mistakes	3.1 (1.45)	4.4 (0.84)

useful for some participants to envision No-Look Notes' layout as a clock (*e.g.*, 'ABC' at 12 o'clock). Participant 1 mentioned this, saying "*a lot of the learning I've done is clock learning anyway.*"

Feedback was mixed on whether the system should repeat a character when it was entered (VoiceOver) or merely give an audible "click" (No-Look Notes). The suggestion of reading the character back but in a different voice (*e.g.*, female) met with a favorable response.

Despite voicing their frustrations with both systems, at the end of the experiment participants were uniformly enthused about learning them. Participants maintained that the experiment had been fun, despite the frustration that was sometimes apparent during testing.

6 Discussion

Our participants' perseverance, despite clear frustration, was impressive. This high tolerance could be evidence that visually impaired users are inured to struggling through poorly designed systems. Tolerance could also be due to the novelty of using a touchscreen or desire that touchscreen phones gain accessibility (participant 1: "*If they were accessible I would buy one*"). Regardless of this tolerance's source, visually impaired users are eager for a touchscreen accessibility solution and willing to put in the time to learn a new system.

While using the systems, users experimented with different ways of holding the device. Some used a single hand, others used thumbs, others set the device on the table and used either one or two hands. Users generally stuck with the same style of input for both No-Look Notes and VoiceOver. The majority of users preferred to leave the device on the table.

Related to device positioning was our decision to test only in a portrait orientation. VoiceOver also allows for a landscape orientation, which would give users more options and, more tangibly, allow VoiceOver to increase the size of its targets by about 25%. While this would likely improve VoiceOver's performance, we feel it is unlikely to have a drastic effect. A slight change in target size would not provide the advantages of full screen use (such as tracing the screen's edges) or the low number of targets present in No-Look Notes, nor would it address the audio feedback issue discussed below.

Some users also displayed a hesitance to enter a character if they weren't sure of the results, hovering their finger above the screen as they meditated on which direction to gesture. There may be a significant psychological difference between a truly risk free 'exploration gesture' and actually entering a character, then deleting it.

6.1 Simple Entry Gestures

Both systems put emphasis on simple input gestures in an attempt to make accidental entry of characters difficult. The split-tapping technique, developed by Slide Rule as a tool for making selections off of lists [11], proved effective at selecting arbitrarily located targets, provided the targets were large enough. VoiceOver's targets were small enough that the participants occasionally moved off of the target while preparing to tap the screen with a second finger.

VoiceOver's double-tapping entry technique was generally more difficult for users, with only one participant preferring it to split-tapping. The key issue was tapping with sufficient speed. If a user was too slow, the system would interpret the tap as a touch and re-locate the system's focus, losing the user's place.

6.2 Text-Awareness

Participants expressed a desire for the ability to easily determine the text they had entered. While some were able to use VoiceOver's cursor system (one even used the cursor to insert missing characters), others found the system difficult to grasp. No-Look Note's simple read-and-spell gesture was usable for all participants, although waiting for the system to spell the word was galling to some.

It is key that users be able to quickly check both their recently entered text and a larger amount of their text. When users get lost or forget their place, they can be forced to erase their text and start anew. A combination of VoiceOver's ability to precisely examine each character combined with a more rapid version of No-Look Note's direct access to spelling may be best.

6.3 Audio Feedback

Audio was a key component of both systems, used for both locating characters and reading text. Users found both systems difficult to understand. The dynamic text to speech engine used in both systems, Flite [18], was particularly disliked.

Though Apple has created a dynamic TTS system for the iPhone, it is inaccessible to developers at the time of this writing. VoiceOver also effectively

cuts developers off from interpreting gestures themselves, making it impossible to use many applications. These design decisions severely restrict developers, preventing them from harnessing the power of gestures and restricting them to button-based systems. This could lead to developers creating unnatural or unusable ports of gesture-based applications, or simply failing to make an accessible version at all.

Nearly all participants had difficulty distinguishing between like-sounding letters (*e.g.*, ‘M’ and ‘N’, ‘C’ and ‘Z’). This led to extreme frustration in VoiceOver’s QWERTY based layout, where like-sounding letters were often adjacent. Even the experimenters found it almost impossible to aurally distinguish certain letter groups. Using this TTS in a noisy environment or using the phone’s built-in speaker will exacerbate the problem, though headphones could help avoid this. A serious portion of VoiceOver’s high error-rate can be attributed to this issue.

No-Look Notes was able to escape this problem with its alphabetical groups, largely keeping like-sounding characters apart. Groups like ‘MNO’ were still usable, as users proved more easily able to locate the correct letter when faced with only a few large targets in alphabetical order.

6.4 Deployment

Our results show that No-Look Notes has a number of advantages over VoiceOver. VoiceOver’s focus-box input technique simply did not translate well to the large number of targets in a soft keyboard. We believe best way to use No-Look Note’s strengths is not to replace VoiceOver, however, but to integrate with it. VoiceOver is used not only for text-entry, but for navigating the entire iPhone interface. Rather than invoking a virtual keyboard when a text field is selected, No-Look Notes could be activated. A simple gesture to return to VoiceOver navigation would complete integration.

7 Future Work

Text-Messaging Evaluation. Many users expressed a desire to send text messages using either system. An extended evaluation of both systems in which users actually send text messages would show how effective each system is at enabling this apparent ‘killer app.’

Visual Feedback. No-Look Notes has no visual feedback to help with text entry or the system’s layout. Users with partial vision would benefit from clear graphics on the touchscreen. Graphics need not even be text — colored or pattern areas could be used to designate pie segments, for example, or the screen could flash when a character was entered. This graphically augmented system could then be tested with both visually impaired and sighted users.

Refine Character-Entry Feedback. Experimenters suggested using different voices for different types of feedback, such as using a different tone of voice to repeat characters as they were entered. Participants were enthusiastic about the idea. This could be an effective way of adding additional contextual information to eyes-free text entry systems. This would also help avoid confusion over whether hearing a character read meant that character had actually been entered or merely been touched.

Extend Character Set. No-Look Notes could be extended to allow entry of numbers or symbols. This could involve adding more targets to the pie menu, or perhaps creating a modal input method of swapping between a ‘letters menu’ and a ‘numbers menu.’ VoiceOver features a numbers/symbols keyboard accessed by entering a ‘more’ button on the soft keyboard.

8 Conclusion

We introduced No-Look Notes, a system for eyes-free mobile text entry using multi-touch input. No-Look Notes was designed to take advantage of other work on accessibility and text entry. We also developed three design principles that are key for eyes-free text entry: Robust Entry Technique, Familiar Layout and Painless Exploration.

We have implemented No-Look Notes on an Apple iPhone. We conducted user trials with visually impaired participants to evaluate both No-Look Notes and Apple’s own VoiceOver system, offering a comparison of the two systems. Our study of VoiceOver is also a first look of what will become the first widely distributed system for eyes-free text entry on the iPhone platform.

Acknowledgements. We are indebted to Anisio Correia and Leigh Cooper, our collaborators at the Center for the Visually Impaired of Atlanta. Rosa Arriaga lent essential guidance and expertise to our statistical analysis. Jeffrey Bigham and Bruce Walker gave invaluable advice, especially concerning our evaluation.

References

1. McGookin, D., Brewster, S., Jiang, W.: Investigating touchscreen accessibility for people with visual impairments. In: Proceedings of the 5th Nordic conference on Human-Computer Interaction: Building Bridges, pp. 298–307 (2008)
2. MacKenzie, I.S., Soukoreff, R.W.: Text Entry for Mobile Computing: Models and Methods, Theory and Practice. *Human-Computer Interaction* 17(2), 147 (2002)
3. Sánchez, J., Aguayo, F.: Mobile Messenger for the Blind. In: Universal Access in Ambient Intelligence Environments, pp. 369-385 (2007)
4. Tinwala, H., MacKenzie, I.S.: Eyes-Free Text Entry on a Touchscreen Phone. In: Proceedings of the IEEE Toronto International Conference Science and Technology for Humanity TIC-STH 2009, pp. 83–89 (2009)

5. Perlin, K.: Quikwriting: continuous stylus-based text entry. In: Proceedings of the 11th annual ACM symposium on User interface software and technology, pp. 215–216 (1998)
6. Plimmer, B., Crossan, A., Brewster, S.A., Blagojevic, R.: Multimodal Collaborative Handwriting Training for Visually-Impaired People. In: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, pp. 393–402 (2008)
7. Yfantidis, G., Evreinov, G.: Adaptive Blind Interaction Technique for Touchscreens. Universal Access in the Information Society 4(4), 328–337 (2006)
8. Speech Enabled Eyes Free Android Applications,
<http://code.google.com/p/eyes-free/>
9. VoiceOver, <http://www.apple.com/accessibility/iphone/vision.html>
10. Mynatt, E.D., Edwards, W.K.: Mapping GUIs to Auditory Interfaces. In: Proceedings of the 5th annual ACM symposium on User Interface Software and Technology, pp. 61–70 (1992)
11. Kane, S.K., Bigham, J.P., Wobbrock, J.O.: Slide Rule: Making Mobile Touch Screens Accessible to Blind people using Multi-Touch Interaction Techniques. In: Proceedings of the 10th international ACM SIGACCESS Conference on Computers and Accessibility, pp. 73–80 (2008)
12. Brewster, S., Chohan, F., Brown, L.: Tactile Feedback for Mobile Interactions. In: Proceedings of the SIGCHI Conference on Human Factors in Computing System, pp. 159–162 (2007)
13. Leung, R., MacLean, K., Bertelsen, M.B., Saubhasik, M.: Evaluation of Haptically Augmented Touchscreen GUI Elements under Cognitive Load. In: Proceedings of the 9th International Conference on Multimodal Interfaces, pp. 374–381 (2007)
14. Kaaresoja, T.: Snap-Crackle-Pop: Tactile Feedback for Mobile Touch Screens. In: Proceedings of Eurohaptics, pp. 565–566 (2006)
15. Yatani, K., Truong, K.N.: SemFeel: A User Interface with Semantic Tactile Feedback for Mobile Touch-screen Devices. In: Proceedings of the ACM Symposium on User Interface Software and Technology, pp. 111–120 (2009)
16. E.161: Arrangement of Digits, Letters and Symbols on Telephones and Other Devices that can be Used for Gaining Access to a Telephone Network,
<http://www.itu.int/rec/T-REC-E.161-200102-I/en>
17. Wobbrock, J.O., Myers, B.A., Kembel, J.A.: EdgeWrite: A Stylus-Based Text Entry Method Designed for High Accuracy and Stability of Motion. In: Proceedings of the 16th annual ACM Symposium on User Interface Software and Technology, pp. 61–70 (2003)
18. Black, A., Lenzo, K.: Flite: A Small Fast Runtime Synthesis Engine. In: 4th ISCA Speech Synthesis Workshop, pp. 157–162 (2007)
19. MacKenzie, I.S., Soukoreff, R.W.: Phrase Sets for Evaluating Text Entry Techniques. In: CHI 2003 Extended Abstracts on Human Factors in Computing Systems, pp. 754–755 (2003)
20. Wigdor, D., Balakrishnan, R.: *TiltText*: Using Tilt for Text Input to Mobile Phones. In: Proceedings of the 16th annual ACM symposium on User interface software and technology, pp. 81–90 (2003)

On the Use of Brain Decoded Signals for Online User Adaptive Gesture Recognition Systems

Kilian Förster¹, Andrea Biasiucci^{2,3}, Ricardo Chavarriaga²,
José del R. Millán², Daniel Roggen¹, and Gerhard Tröster¹

¹ ETH Zurich, IFE, Wearable Computing Lab
CH-8092 Zurich, Switzerland

² EPFL, CNBI, Center for Neuroprosthetics,
CH-1015 Lausanne, Switzerland

³ University of Genova

Department of Informatics, Systems and Telematics (DIST),
16126 Genova, Italy

{foerster,roggen,troester}@ife.ee.ethz.ch,
{andrea.biasiucci,ricardo.chavarriaga,jose.millan}@epfl.ch

Abstract. Activity and context recognition in pervasive and wearable computing ought to continuously adapt to changes typical of open-ended scenarios, such as changing users, sensor characteristics, user expectations, or user motor patterns due to learning or aging. System performance inherently relates to the user’s perception of the system behavior. Thus, the user should be guiding the adaptation process. This should be automatic, transparent, and unconscious.

We capitalize on advances in electroencephalography (EEG) signal processing that allow for error related potentials (ErrP) recognition. ErrP are emitted when a human observes an unexpected behavior in a system. We propose and evaluate a hand gesture recognition system from wearable motion sensors that adapts online by taking advantage of ErrP. Thus the gesture recognition system becomes self-aware of its performance, and can self-improve through re-occurring detection of ErrP signals.

Results show that our adaptation technique can improve the accuracy of a user independent gesture recognition system by 13.9% when ErrP recognition is perfect. When ErrP recognition errors are factored in, recognition accuracy increases by 4.9%. We characterize the boundary conditions of ErrP recognition guaranteeing beneficial adaptation. The adaptive algorithms are applicable to other forms of activity recognition, and can also use explicit user feedback rather than ErrP.

1 Introduction

Human activity and gesture recognition from body worn motion sensors using machine learning techniques [1] enables activity based computing [2].

Motivation. Activity recognition systems are trained in a user-independent manner for ‘out of the box’ operation. Training data is collected from multiple

subjects to build generic statistical activity models. Exhaustive data collection is time consuming and may not be practical. The recognition of simple activities may already be difficult in a user independent manner [3]. As activities get more complex, this becomes a major challenge. Some highly complex gesture recognition systems weren't even trained for user independence [4].

User specific models usually perform better than user independent models but are less able to generalize to new subjects [5,6,3]. They are trained on the target user to reflect individual characteristics. This individual training phase may not be practical when deploying a system.

User independent and user specific systems are trained once at design time, respectively during first use, and remain static throughout operation. Thus they are not able to adapt to so far unseen situations typical of open-ended scenarios. These systems also have no knowledge about their instantaneous performance, as it was characterized at training time on a specific dataset. Therefore no action can be taken if runtime performance drops. This can be important for example in critical applications, where stopping the context-aware system may be better than letting it operate with reduced performance.

Contribution. Activity and context recognition in pervasive and wearable computing ought to continuously adapt to changes typical of real-world applications, such as a new user of the system, changing sensor characteristics, changing user expectations, or changing user motor patterns due to learning or aging. System performance inherently relates to the user's perception of the system behavior. Thus, the user should be guiding the adaptation process. This should be automatic, transparent, and unconscious.

In order to guide adaptation according to the user's run-time expectation, a feedback signal is required. We capitalize on advances in electroencephalography (EEG) signal processing that allow for error related potentials (ErrP) recognition. ErrP occur when a human observes an unexpected behavior in a system [7,8,9]. We propose and evaluate a hand gesture recognition system from wearable motion sensors that adapts online by taking advantage of ErrP. Essentially the activity recognition system turns into an autonomous system with performance self-awareness and self-improvement capabilities.

In this work, we focus on user specific adaptation from a user-independent model through ErrP signal occurrences. Specific contributions include:

- An experimental setup (gesture based HCI scenario) that allows the joint investigation of activity recognition, ErrP detection, and the combination of both into an autonomously adaptive activity recognition system.
- A dataset of EEG signals, electro-oculography (EOG), hand acceleration and electro-myography (EMG), with 18'000+ gesture instances on 7 subjects.
- The baseline ErrP detection and non-adaptive gesture recognition accuracy.
- A method to estimate instantaneous recognition performance from ErrP.
- A comparative analysis of three strategies to adapt the gesture recognition system to a specific user from a user independent model based on ErrP.

Paper content. In section 2 we review adaptive approaches applied to activity recognition and explain the nature of ErrP. In section 3 we describe the experimental setup we use to investigate adaptive activity recognition driven by ErrP. In section 4 we present the ErrP detection results. In section 5 we show how instantaneous system performance can be derived from ErrP signals. We comparatively evaluate user adaptation strategies driven by ErrP signals in section 6. We discuss results in section 7 and conclude in section 8.

2 State of the Art

Adaptation strategies and limitations. Adaptive techniques can improve the performance for individual users without affecting generalization capability. A user-independent model can adapt to a specific user during a short calibration phase. This has been investigated in handwriting [10] and speech [11] recognition, where it remains a major research topic [12]. Similar approaches were proposed in gesture recognition [13]. Calibration-based approaches are time consuming with large number of activities and activity models remain static after calibration. In dynamic model selection a pre-existing model that best corresponds to the current user or his environment is selected at run-time. This has been applied in speech processing [14]. In activity recognition it has been used to adapt to the user's on-body sensor placement preferences, by selecting models corresponding to the automatically detected sensor location [15]. Such approaches require extensive training data to build multiple models. Other adaptation techniques rely on the unsupervised tracking of clusters of activities in the feature space [16]. While devised for sensor placement adaptation, similar principles may apply to user adaptation. Such approaches can adapt at run-time because they rely on underlying data structure properties. However they do not guarantee to adapt in a way that reflects the user's perception of system performance.

Current adaptation strategies do not take into account the user's perception of the system's behavior. Guiding adaptation according to the user's run-time expectation requires a feedback signal. Explicit interaction may provide this feedback, such as a button that is pressed when the behavior of the context-aware system is not satisfactory. In the vision of wearable and pervasive computing, however, feedback should be transparent: automatic and unconscious.

Brain signals related to unexpected action perception. Several studies have suggested the existence of a neural system responsible for error processing [17]. Specifically, stereotypical electrophysiological signals have been consistently reported to appear as a response to erroneous actions [18] or unexpected action outcomes [9]. These signals, —termed *Error-related negativity* (ERN) and *Feedback-related negativity* (FRN)— are characterized by a negative deflection of the EEG signals in fronto-central areas of the scalp, followed by a centro-parietal positive peak. Typical signal latencies are 50 to 100ms in the case ERNs and around 250ms for FRNs. Neurophysiological studies have provided evidence of error-based learning. Specifically, it has been suggested that these signals reflect

conscious error processing; post-error adjustment of response strategies [18], and reward-based adaptive behavior [9].

Moreover, research on Brain-Computer Interfaces (BCI) has shown that it is possible to recognize EEG error-related signals (ErrP) on single trials above random levels [19][20][21]. Based on this fact, these signals have been proposed to be used to correct erroneous motor action in speed-response human-computer interaction [21], as well as to increase the information transfer rate of EEG-based BCI systems [19]. Experimental measures taken over different time periods (up to two years) show that these potentials are stable over time, despite the delay between recordings. Current protocols for EEG signal analysis require motionless subjects to avoid that EMG signals (1-30mV) from muscle activity contaminate the subtler EEG signals (10-100 μ V) [22]. In order to use EEG system in naturalistic settings, however, researchers now start to investigate limited subject mobility.

3 ErrP-Based Adaptive Gesture Recognition Scenario

We investigate the use of ErrP to guide the adaptation of a gesture recognition system in an HCI scenario. This scenario is based on a game to maintain the user's involvement during experimental sessions [23][19]. It is designed so that a large number of gesture instances can be acquired in a comparatively short amount of time. It allows movements of the user's arm with limited amplitude, to investigate EEG signal analysis in more realistic situations than state of the art EEG protocols. EEG signal and hand acceleration from a wearable sensor is recorded during the scenario to assess adaptation strategies in offline simulations.

Gesture-controlled computer game. The subjects played a computerized version of a “memory game” consisting of 8 image pairs (fig. II). The 16 images are randomly distributed in a four by four matrix and hidden behind question marks. The subjects have to find identical pairs of images, which are then removed from the screen. If two images are flipped they are hidden again before a new image can be selected. The game is finished when all image pairs were correctly found.

The game input interface is based on five hand gestures. Left, right, up and down hand movements shift the image selection cursor in the corresponding direction. Each directional gestures starts and ends at a central home position. Flipping an image is controlled by closing and opening the hand.

Measurement setup. The online recognition of the gestures is based on light barriers and a reed switch. This ensures accurate gesture recognition for the collection of a reference dataset. Three horizontal and three vertical infrared light barriers detect the hand position (see fig. II). The closing gesture is detected from a reed switch on the subjects hand activated by a magnet on the subjects fingers.

A tri-axial acceleration sensor at the subjects fingertips records the motion of the hand for offline acceleration-based gesture recognition. The acceleration sensor is sampled at 64 Hz and connected via USB to the experiment computer. This computer also ran the memory game. Another computer recorded EEG, as

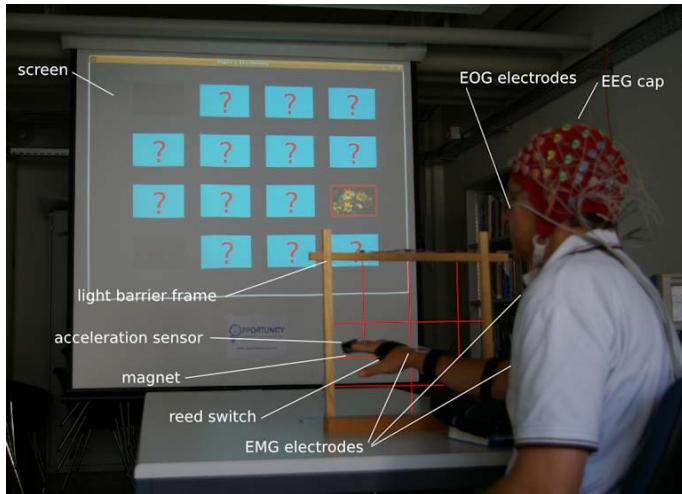


Fig. 1. The computer game is presented on the screen; the light-barrier frame, magnet and reed switches capture game control gestures; the acceleration sensor, EMG electrodes (right wrist, biceps and shoulder), EOG electrodes, and EEG electrode cap stream data to a PC for recording and offline analysis

well as arm EMG and eye movements using EOG with the Biosemi ActiveTwo system with active electrodes. EMG and EOG allow for motion artifacts cancellation by adaptive filtering. Both computers were interconnected to ensure a synchronized data recording using a shared data line.

Experimental protocol. Seven healthy male subjects aged 25 to 47 participated. For each subject we recorded 14 sessions with a duration of three to five minutes. One session corresponds to one memory game. Between recording sessions the subjects could rest for one to two minutes. We recorded more than 2700 hand gestures per subject. The experiment lasted about two hours per subject including setup.

In each session we randomly artificially induced between 5% and 33% of gesture recognition errors to provoke ErrP events. In an error case the game selects a random command instead of the user command. For example if the subject closes his hand to turn a card, the card is not turned but instead the cursor is moved in a random direction.

4 EEG-ErrP Single-Trial Recognition

Following previous studies [23,19], we perform classification using the time signal of electrodes FCz and Cz as input features for a Bayesian filter [24], since EEG ErrP are characterized by a fronto-central distribution along the midline. EEG potentials were spatially filtered by subtracting from each electrode the average

potential (i.e. the common average reference) at each time step to suppress average brain activity and keep the information from the local sources below each electrode. A 1–10-Hz bandpass filter was applied as ErrPs are relatively slow cortical potentials [25]. EEG signals were subsampled to 64Hz before classification, which is based on temporal features. The input vector for the classifier (described below) is composed by the time samples on electrodes FCz and Cz within a fixed time window after the feedback onset.

At each sampling time step, the Bayesian filter estimates the state probabilities according to the observations and the previous state estimations. In this case, we have discrete observations of a continuous EEG signal and we want to find the state for the action shown on the screen, i.e. an erroneous or correct movement.

To build the Bayesian filter, we define two possible states at each time t : $S_t = 1$ for erroneous recognition, and $S_t = 0$ for correct recognition. At each sampling time step t observations O_t are given by a vector with components FCz and Cz corresponding to the electrodes of the same name: $O_t = [FCz_t, Cz_t]$. Observations and states from time zero to T are respectively noted $O_{0:T}$ and $S_{0:T}$.

A transition model is defined by a first order Markov hypothesis for states over time: $P(S_t|S_{0:t-1}) = P(S_t|S_{t-1})$ for $t = 0 \dots T$. Since the state during a single trial doesn't change, the transition model corresponds to the identity matrix: $P(S_t|S_{t-1}) = 1$ if $S_t = S_{t-1}$ and zero otherwise.

The sensor model is given by the probability distribution $P(O_t|S_t)$ which predicts observations given the state. Then the decomposition of the joint probability is given by:

$$P(S_{0:T}O_{0:T}) = P(S_0)P(O_0|S_0) \prod_{t=1}^T (P(S_t|S_{t-1})P(O_t|S_t)) \quad (1)$$

The classification consists in estimating $P(S_t|O_{0:t})$, i.e. the probability of the state (error or correct) knowing the observations (EEG activity). It can be obtained in a recurrent manner; first, a *prediction* (2) of the state is done based on the transition model and then, second, the state *estimation* (3) is computed based on the sensor model.

$$P(S_t|O_{0:t-1}) = \sum_{S_{t-1}} (P(S_t|S_{t-1})P(S_{t-1}|O_{0:t-1})) \quad (2)$$

$$P(S_t|O_{0:t}) \propto P(O_t|S_t)P(S_t|O_{0:t-1}) \quad (3)$$

Given the identity transition matrix, the *prediction–estimation* recurrent calculus is simplified:

$$P(S_t = 1 | O_{1:t}) \propto P(O_t|S_t)P(S_{t-1} = 1 | O_{1:t-1}) \quad (4)$$

And correspondingly for $P(S_t = 0 | O_{1:t})$. Being Q_t be the quotient of the probabilities for both states, an erroneous trial is detected when $\ln(Q_t)$ is positive, where $\ln(Q_t) = \ln(Q_{t-1}) + \ln(P(O_t|S_t = 1)) - \ln(P(O_t|S_t = 0))$.

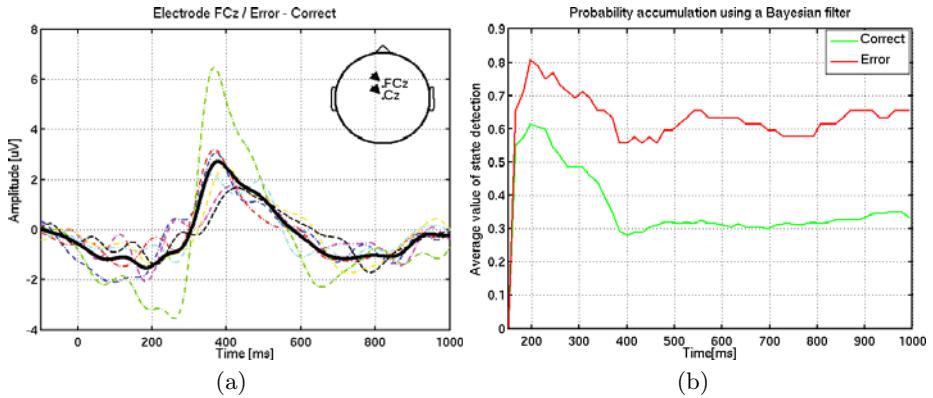


Fig. 2. (a) Grand average error-related potential on the FCz electrode, Error minus Correct condition, (*thick line*); individual subject averages are shown with *thin dashed lines*. Time ($t=0$) is measured from the feedback onset. Electrodes positions are shown in the scalp plot. (b) Average value of the state detection for correct and erroneous trials of session 1 on subject V.

Estimations from both channels are combined using a naive fusion,

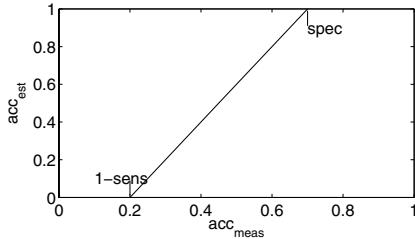
$$P(O_t|S_t) = P(FCz_t|S_t)P(Cz_t|S_t) \quad (5)$$

The sensor model $P(O_t|S_t)$ is defined by a Gaussian distribution with a mean μ_t and a variance σ_t^2 , these parameter were estimated using the training dataset. Having two input channels and two possible states, we have four Gaussian distributions at each time t , and eight parameters to identify. This approach updates the estimated state probability as new samples are available, an example of the average estimated probability at different time points. Fig. 2 shows the average EEG activity in channel FCz (error-minus correct condition), as well as the average state estimation for both classes at different time points.

We divided the recorded dataset into seven folds of EEG activity for each subject; every single fold corresponds to two consecutive memory games. The classifier was trained using a Leave-One-Fold-Out Cross Validation, i.e. training with 6 folds, and testing on the 7th, then averaging results for all folds. We consider the activity of electrodes in the [150, 1000] ms time windows after the feedback presentation, estimating the state probabilities according to the observations and taking a choice at the time instant that maximizes State Recognition. Table I shows classification results for the described technique; we should stress the fact that the ErrP recognition can be affected by EMG artifact contamination due to subject movements. This indeed a major challenge in the integration of EEG activity in pervasive applications. Different filtering techniques can be applied for reducing such contamination [26], and its use will be object of further study.

Table 1. Sensitivity and specificity for the seven subjects using Leave-One-Fold-Out Cross Validation

	Subjects							Average
	I	II	III	IV	V	VI	VII	
<i>Sensitivity</i>	0.74	0.56	0.60	0.63	0.57	0.65	0.58	0.62
<i>Specificity</i>	0.48	0.71	0.73	0.59	0.75	0.65	0.63	0.65

**Fig. 3.** The estimated accuracy vs. the measured accuracy for a specific specificity and sensitivity level

5 Performance Self-awareness through ErrP Detection

A run-time system performance measure can be obtained from the detection of ErrP events. We define N_{err} the number of ErrP events and N_{gest} the number of executed gestures during a period of operation. The ErrP detection specificity and sensitivity is $spec_{ErrP}$ and $sens_{ErrP}$, characterized during system training. We assume that sensitivity and specificity is stationary throughout operation (i.e. no fluctuations of $spec_{ErrP}$ and $sens_{ErrP}$). In this case we can estimate the true accuracy of the gesture recognition as follows:

$$acc_{est} = \frac{acc_{meas} - 1 + sens_{ErrP}}{spec_{ErrP} + sens_{ErrP} - 1} = \frac{1 - \frac{N_{err}}{N_{gest}} - 1 + sens_{ErrP}}{spec_{ErrP} + sens_{ErrP} - 1} \quad (6)$$

The specificity and sensitivity of the ErrP detection control slope and the offset of the dependency between acc_{meas} , the measured accuracy based on the ErrP signals, and acc_{est} , the estimated true accuracy of the system, as depicted in figure 3. As $spec_{ErrP}$ and $sens_{ErrP}$ get lower, slight errors in the measurement of acc_{meas} will have more effect on the estimated true accuracy (due to the steeper line slope in the figure).

6 ErrP-Based Gesture Recognition Adaptation

ErrP signals indicate when the action taken by a system is erroneous (see table II), and thus whether a gesture was wrongly recognized in our game scenario. The open questions are what are the adaptive strategies suitable to incorporate the information provided by ErrP into a gesture recognition system.

User-specific adaptation scenario. We consider an adaptation scenario where a user independent gesture recognition system is adapted to a specific user through ErrP occurrences. The gesture recognition system is trained in a user independent manner in the HCI scenario (i.e. it uses a user independent classifier C_{init}). The system is then given to a so far unseen user. Each gesture performed by this user is classified by the gesture recognition system. EEG analysis indicates whether the action taken by the computer game, and thus the classification of the gesture, was correct or wrong. We adapt the gesture classification system through online learning (see below).

Gesture classification. We distinguish the five game control gestures based on the hand acceleration. We segment the signal using the gesture-start and gesture-end signal provided by the light-barrier frame. During training of classifiers, the ground truth label of gesture instances was provided by the light-barrier frame. We did no dataset cleaning or outlier removal as this would not be possible in a real application of this kind.

For each gesture, we calculate the following acceleration features on three windows (full gesture, first half and second half of the gesture): mean, standard deviation, minimum, maximum and energy. We do this on the three axes of the acceleration signal as well as on its magnitude. In addition the correlation for each axes pair xy, xz and yz is calculated. This yields 63 features. We perform a probabilistic feature selection [27] combined with a scatter search [28] to select a feature subset. This yields 6 features: the mean on y axis, the mean and minimum on magnitude, the mean on first half of z axis, the standard deviation on first half of x axis and the mean on second half of y axis.

We classify the gestures with the following classifiers: Naive Bayes [29], Bayes Networks [30] and k Nearest Neighbor (kNN) [31] implemented in the Weka Machine Learning Project [32]. For the kNN classifier we chose $k = 13$ as it gave sufficiently good results for all subjects. A higher value for k would also increase the minimum number of training instances. We use a batch approach [33] for on-line learning.

Strategies to exploit EEG. The absence of ErrP indicates that the classification result of the system is correct. We assume this result is the ground truth class label of the gesture. The presence of ErrP indicates a wrong classification, but does not provide indication of the class label. Therefore during operation we can collect labeled user specific samples (those where ErrP was not detected). We investigate three strategies to create user adapted classifiers. Essentially all strategies start from a user independent classifier C_{init} and operate by collecting a user specific training set S_x . They then train a user adapted classifier C_x on this set. The strategies differ in the way the training set is collected:

1. **AD 1 “Incremental knowledge integration”:** Starting from a user independent training set $S_1 = S_{init}$, the user adapted training set S_1 grows by including a new (user specific) gesture instance whenever the gesture performed by the user is recognized and no ErrP is detected (see algorithm II).

Algorithm 1. Adaptation strategy 1

1. Initialize dataset S_1 with n_{init} subject independent instances
 2. Train classifier C_1 on the subject independent initial training set S_1
 3. **for** each user specific instance i **do**
 4. classify instance i using C_1 to class c
 5. **if** no error detected **then**
 6. add instance together with label c to S_1
 7. retrain C_1 on the new S_1
 8. **end if**
 9. **end for**
-

After a new gesture is collected a classifier C_1 is trained on S_1 and replaces C_{init} .

2. **User specific batch training:** a batch of user specific training samples S_{2x} is collected whenever a gesture performed by the user is recognized and no ErrP is detected. There are two training sample collection variants:
 - (a) **AD 2a:** the user independent classifier C_{init} classifies the gestures (see algorithm 2)
 - (b) **AD 2b:** the classifier created with adaptation strategy 1 (AD1) classifies the gestures (see algorithm 3)
-

Algorithm 2. Adaptation strategy 2a

1. Initialize dataset S_{init} with n_{init} subject independent instances
 2. Initialize dataset S_{2a} to the empty set
 3. Train classifier C_{init} on the subject independent initial training set S_{init}
 4. **for** each user specific instance i **do**
 5. classify instance i using C_{init} to class c
 6. **if** no error detected **then**
 7. add instance i together with label c to S_{2a}
 8. **if** S_{2a} contains sufficient instances **then**
 9. train classifier C_{2a} on S_{2a}
 10. **end if**
 11. **end if**
 12. **end for**
-

To train the user independent classifier C_{init} we combine the data of all subjects, leaving out the subject we want to adapt to. From this combined dataset we select randomly n_{init} instances which are used for the training. The data of the left out subject is split into an *adaptation set* and a *test set*. The *adaptation set* contains 2250 instances while the *test set* contains 500 instances. We purposely preserve the timely order of the data instances in the *adaptation set* to simulate the adaptation as close to reality as possible. During operation, the instances in the *adaptation set* are iteratively presented to the system for classification and adaptation.

Algorithm 3. Adaptation strategy 2b

1. Initialize dataset S_1 with n_{init} subject independent instances
 2. Initialize dataset S_{2b} to the empty set
 3. Train classifier C_1 on the subject independent initial training set S_1
 4. **for** each user specific instance i **do**
 5. classify instance i using C_1 to class c
 6. **if** no error detected **then**
 7. add instance i together with label c to S_1
 8. add instance i together with label c to S_{2b}
 9. retrain C_1 on the new S_1
 10. **if** S_{2b} contains sufficient instances **then**
 11. train classifier C_{2b} on S_{2b}
 12. **end if**
 13. **end if**
 14. **end for**
-

6.1 Adaptation Assuming Perfect ErrP Detection

For the following simulation we assume a perfect ErrP detection with a sensitivity and a specificity of 1. In figure 4 we show the evolution of the classification accuracies for the different adaptation strategies over time. Due to space constraints we limit ourselves to the Bayes Network classifier, the other classification methods show the same trends. Each plot shows adaptation from a different number of initial subject independent training instances n_{init} . Every 45 iterations the accuracy of each classifier with respect to the subject dependent *test set* is given. Two baselines represent the performance of a non-adaptive subject independent and subject dependent classifier. The subject independent classifier is trained on the initial training set with n_{init} instances. The subject dependent classifier is trained on the *adaptation set*. It indicates the upper bound an adapted user specific classifier can achieve if all the user specific information would be available.

The classifiers built based on adaptation strategy 1 show a benefit over the subject independent baseline. The larger n_{init} the smaller the gain in accuracy.

The classifiers built based on adaptation strategies 2a and 2b outperform the subject independent baseline and also adaptation strategy 1 for higher n_{init} . As the number n_{init} of initial subject independent training instances increases, the 2a and 2b adapted classifiers reach a higher accuracy.

To further the benefit of the additional ErrP information we also show a baseline for adaptation strategy 1 in a setting where every classification result is assumed to be correct - simulating the absence of ErrP. Without ErrP the adaptation leads to a classifier performing worse or only slightly better than the user independent classifier in this setting.

The marker in the plots show the points where the 1, 2a and 2b adapted classifiers reach 99.9% of the accuracy they achieve after all iterations. This indicates how many iterations are sufficient to build a good adapted classifier.

The results averaged over all subjects and over five simulation runs from different random seeds are listed in table 2. In all cases at least one of the user

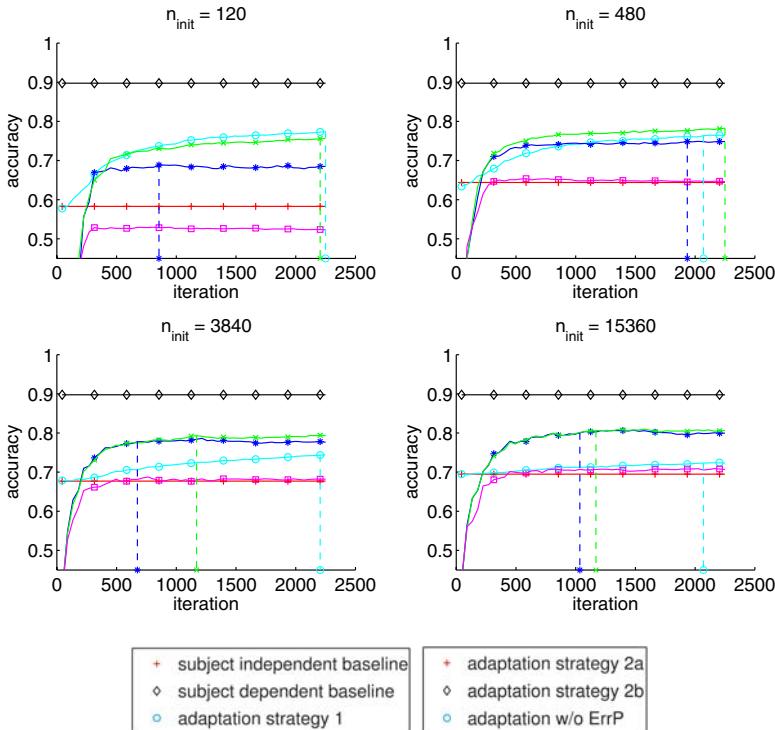


Fig. 4. Accuracies of the Bayes Network built based on adaptation strategies 1, 2a and 2b over time for four different numbers of instances n_{init} in the initial training set. At each iteration the adapted classifiers are tested on the subject dependent *test set*. The dashed lines mark the point where 99.9% of the accuracy after all iterations is reached.

adaptation strategies leads to a classifier outperforming the subject independent classifier. The best adaptation is achieved with strategy 2b in most cases. The average improvement achieved by the best adaptation strategy over the subject independent case is 12.1% for the Naive Bayes, 13.9% for the Bayes Network and 12.2% for the kNN.

The Symbols (+, *) next to the recognition accuracy indicate for how many of the subjects the increase in performance provided by the adaptation strategy is statistically significant compared to the subject independent accuracy (T-Test, $p < 0.05$, null hypothesis is that the performance after adaptation is identical to the subject independent accuracy).

6.2 Adaptation Using Experimental ErrP Detection Performance

We present the results obtained from the experimentally measured ErrP detection accuracy. ErrP detection is challenging in this setup and the average sensitivity is 0.65, the average specificity 0.62, over all subjects and sessions,

Table 2. Perfect ErrP detection: Accuracies achieved with the subject dependent (SD), subject independent (SID), 1 adapted (AD 1), 2a adapted (AD 2a) and 2b adapted (AD 2b) classifiers in %. The values in brackets give the number of iterations until 99.9% of the final accuracy is reached. (+: significant increase in accuracy for at least 3 out of 7 subjects, *: significant increase in accuracy for at least 5 out of 7 subjects).

	n_{init}	SID	AD 1	AD 2a	AD 2b
Naive Bayes	120	56.6	69.3* (1845)	69.6* (2205)	70.8* (1710)
	480	59.1	67.8+ (2205)	70.4* (1485)	72.2* (1755)
	SD = 75.7	3840	60.7	62.8+ (2205)	71.1* (360)
	15360	60.7	60.8+ (180)	71.3* (360)	71.1* (450)
Bayes Network	120	58.3	77.4* (2250)	68.5* (855)	75.6* (2205)
	480	64.4	76.4* (2070)	74.9* (1935)	78.2* (2250)
	SD = 89.7	3840	67.7	74.4* (2205)	77.6* (675)
	15360	69.5	72.3* (2070)	80.0* (1035)	80.6* (1170)
kNN k = 13	120	57.7	74.2* (2160)	70.7* (2160)	76.9* (2025)
	480	67.6	73.9* (315)	75.9* (810)	79.9* (2205)
	SD = 92.2	3840	73.7	82.2* (2205)	71.6* (270)
	15360	76.4	81.6* (2250)	84.1* (2250)	85.2* (2250)

which is only slightly above chance. We repeated the simulations from section 6.1 taking into account the inaccuracies in the error detection. The simulation results are listed in table 3.

For all values of n_{init} at least one adapted classifier performs better than the subject independent classifier. The gain is however marginal in several cases and certain adapted classifier perform worse than the subject independent one. Adaptation strategy 2a performs best for the Naive Bayes classifier while strategy 2b is more appropriate for the Bayes Network and the kNN.

The average improvement achieved by the best adapted classifier over the subject independent classifier is 2.9% for the Naive Bayes, 3.7% for the Bayes Network and 4.9% for the kNN.

6.3 Influence of the ErrP Detection Accuracy on the Adaptation

The ErrP recognition performance is a key parameter for a successful adaptation. A perfect ErrP recognition leads to a performance of the user adapted classifiers higher than the user independent classifier. With the ErrP performance experimentally achieved in our setup the improvement is comparatively lower. Typical EEG ErrP recognition algorithms can be adjusted towards increased specificity or sensitivity following a ROC curve. By understanding the range of ErrP recognition sensitivity and specificity values where the adaptation of the gesture recognition shows benefit it becomes possible to adjust the ErrP recognition parameters along the ROC curve to ensure a benefit.

We consider the adaptation to be beneficial when the user adapted classifier performs significantly better for at least 3 out of 7 subjects. We consider the 2a and 2b adaptation strategies only with $n_{init} = 480$ instances to reduce the computation effort.

Table 3. Experimental ErrP detection: Accuracies achieved with the subject dependent (SD), subject independent (SID), 1 adapted (AD 1), 2a adapted (AD 2a) and 2b adapted (AD 2b) classifiers in %. The values in brackets give the number of iterations until 99.9% of the final accuracy is reached. (+: significant increase in accuracy for at least 3 out of 7 subjects, *: significant increase in accuracy for at least 5 out of 7 subjects).

	n_{init}	SID	AD 1	AD 2a	AD 2b
Naive Bayes	120	56.6	58.3 (990)	59.6 (900)	59.1 (1080)
	480	59.1	59.0 (1440)	62.3 ⁺ (360)	60.5 ⁺ (1080)
	SD = 75.7	60.7	58.7 (45)	63.4 ⁺ (270)	62.6 ⁺ (315)
	15360	60.7	59.3 (45)	63.6 [*] (360)	63.4 [*] (315)
Bayes Network	120	58.3	62.9 (2160)	59.3 (315)	61.3 (2205)
	480	64.4	67.3 ⁺ (1890)	66.5 (720)	67.8 ⁺ (1485)
	SD = 89.7	67.7	69.5 (2205)	70.3 ⁺ (675)	70.6 ⁺ (2250)
	15360	69.5	70.1 (1755)	73.1 ⁺ (2070)	72.5 ⁺ (855)
kNN k = 13	120	57.7	62.5 ⁺ (1890)	60.5 (1080)	63.6 ⁺ (1800)
	480	67.6	71.1 ⁺ (315)	71.4 ⁺ (1890)	71.0 ⁺ (765)
	SD = 92.2	73.7	78.5 [*] (2070)	77.0 [*] (2160)	79.1 [*] (2250)
	15360	76.4	79.6 [*] (2205)	79.6 [*] (1890)	81.1 [*] (2250)

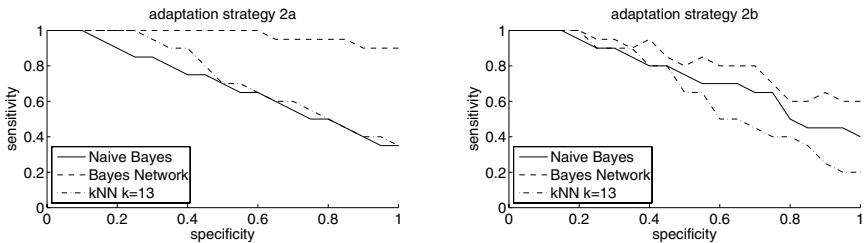


Fig. 5. The curves for each classification method give the ErrP detection performance for which the adapted classifier improves significantly for at least three subjects over the subject independent classifier. For every point to the right or above these curves a benefit can be expected when adapting based on a ErrP detection system offering this performance.

Figure 5 shows the decision line between sensitivity and specificity pairs beyond which the 2a or 2b adaptation strategy is beneficial. Any point right or above the line describes an ErrP detection sensitivity or specificity that allows for beneficial adaptation. For strategy 2a the Naive Bayes and the kNN are quite robust while the Bayes Network shows the least robustness with respect to ErrP recognition performance. With strategy 2b the trends are similar, the Bayes Network is more comparable to the other classification algorithms, though. All algorithms are more sensitive to a low sensitivity than to a low specificity.

7 Discussion

In this work we assess, for the first time, single-trial recognition of EEG error-related potentials in a complex, realistic task. This contrasts with previously reported experiments where these signals were studied using very simple stimuli,

and subjects movements were restricted to minimize motion-related artifacts in the EEG signal [23][19][24]. The difference in the experimental protocol - i.e. subject moving during the recording, complex visual feedback, different cognitive demand of the experimental task -, together with the intrinsic variability, noise, and non-stationarities of brain signals, may explain the low classification accuracies obtained in the current study. Nevertheless, it should be considered that it is not possible to achieve perfect decoding of brain generated signals due to several reasons (i.e. low signal-to-noise ratio, EEG non-stationarity, muscular contamination). Indeed, classification performance for ErrP recognition in much simpler, controlled experiments is approximately 80% for both classes [23][19]. Therefore our results are encouraging.

Despite the low ErrP recognition performance it was still possible to use this additional source of information to successfully adapt the gesture recognition system towards a specific user. The gain in accuracy achieved by the adaptation is depending on the ErrP recognition. The better the ErrP recognition performs the more improvement in the gesture classification can be expected. This is true for all classification methods we investigated.

As we rely on a subject independent gesture recognition system as a basis for our adaptation it is indispensable that this initial system reaches a certain recognition performance. If for example the initial system could not at all recognize one specific class it would not be possible to build a new subject dependent classifier, as this class would be missing in the collected subject dependent data.

In our experimental setup for data collection we assume that the subject intention is correctly captured by the gesture recognition. There might still be cases where the subject performs a wrong gesture by mistake. This mistake might also be reflected in the brain signal as an error. As we do not capture the users intention directly we can not assess the influence of user mistakes.

The gesture recognition errors are added artificially and randomly so that the user can not adapt to it to improve the gesture recognition. Therefore the simulated improvements of the gesture recognition are independent of potential user adaptation.

One can argue that the simulations based on offline data are not meaningful for a live system as other effects, like user adaptation, may come into play. In a live system it is very difficult to investigate all parameters essential for such an adaptation scenario, though.

Even though the ErrP adapted classifier reaches a promising accuracy gain, the performance of a classifier trained in a pure subject dependent manner is still not reached. This can be explained by the fact that the adaptation only uses the instances which are correctly classified by the subject independent classifier. The instances which are too different and therefore not covered by the subject independent model are excluded. These excluded instances potentially contain information important for building a good user dependent model.

In the adaptation schemes we propose we do not make use of confidence values which could be provided by the ErrP detection. These confidence values may be used to weight the instances for the adaptation process.

The online learning we made use of was based on batch learning. Batch learning in general puts higher memory requirements on a system, compared to incremental learning, as all training instances in the batch have to be stored. Especially for wearable systems with limited memory and processing capabilities incremental learning should be considered for online learning and adaptation.

8 Conclusion

We have investigated strategies for user adaptation within a gesture based HCI scenario making use of additional information provided by EEG ErrP analysis. To our knowledge, this is the first attempt to use brain signals related to the perception of errors for the improvement of activity recognition systems. Simulations of a perfect decoding of such signals show that theoretically the recognition accuracy can be increased by up to 13.9% over the user independent classifier. Using single-trial recognition of actual EEG data recorded during the gesture based HCI experiment, the accuracy increase for the adapted gesture recognition reached 4.9% in the best case. This shows that brain signals (i.e. EEG) generated during real human-computer interaction provide information that can be integrated into the activity recognition chain so as to improve its performance.

EEG-based user adaptation remains unlikely in real-world scenarios in the near future given the current state of the sensing technology, its sensitivity to motion artifacts, and the desire for invisible wearables. Miniaturized sensing platforms may become available [34], however there are also many professional occupations that require to wear a helmet or head protection gear (e.g. firefighters, soldiers, surgeons, pilots). In this case the integration of EEG within the head apparel can be envisioned. Since these are usually high stakes professions, a continuous self-monitoring of wearable system performance and its improvement over time may be strong factors supporting the inclusion of such technology. In general there are many potential applications, ranging from disabled people to entertainment [35], which could benefit from “human in the loop” strategies.

An immediate outcome of this work, however, is the comparative evaluation of user adaptation strategies, that are applicable to other forms of user feedback. For instance, a button integrated in a smart shirt or a user interface element could be used to signal a non-desired behavior triggering system adaptation.

Besides using the strategies presented here to adapt a generic classifier to a specific user, they may also be used to deal with changing user preferences or non stationarities, either using implicit EEG-based feedback, or explicit feedback.

In future work we plan to use the recorded EMG and EOG to filter out muscular artifacts that may contaminate the signals used for classification. This may lead to an increase in recognition performance and a higher robustness to contaminations.

To improve the user adaption we further plan to investigate how online learning methods can make additional use of user specific instances which were wrongly classified. Those instances may add valuable information to the adaptation process.

Acknowledgements

The project OPPORTUNITY acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 225938.

References

1. Ward, J.A.: Activity Monitoring: Continuous Recognition and Performance Evaluation. PhD thesis, ETH Zurich, Nr. 16520 (2006)
2. Davies, N., Siewiorek, D.P., Sukthankar, R.: Special issue: Activity-based computing. *IEEE Pervasive Computing* 7(2), 20–21 (2008)
3. Ravi, N., Dandekar, N., Mysore, P., Littman, M.L.: Activity recognition from accelerometer data. *American Association for Artificial Intelligence* (2005)
4. Stiefmeier, T., Roggen, D., Ogris, G., Lukowicz, P., Tröster, G.: Wearable activity tracking in car manufacturing. *IEEE Pervasive Computing Magazine* 7(2), 42–50 (2008)
5. Bao, L., Intille, S.S.: Activity recognition from user-annotated acceleration data. In: *Pervasive Computing: Proc. of the 2nd Int Conference*, pp. 1–17 (2004)
6. Lester, J., Choudhury, T., Borriello, G.: A practical approach to recognizing physical activities. In: Fishkin, K.P., Schiele, B., Nixon, P., Quigley, A. (eds.) *PERVASIVE 2006. LNCS*, vol. 3968, pp. 1–16. Springer, Heidelberg (2006)
7. Nieuwenhuis, S., Ridderinkhof, K.R., Blom, J., Band, G.P., Kok, A.: Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology* 38(5), 752–760 (2001)
8. Yasuda, A., Sato, A., Miyawaki, K., Kumano, H., Kuboki, T.: Error-related negativity reflects detection of negative reward prediction error. *Neuroreport* 15(16), 2561–2565 (2004)
9. Frank, M.J., Woroch, B.S., Curran, T.: Error-related negativity predicts reinforcement learning and conflict biases. *Neuron* 47(4), 495–501 (2005)
10. Santosh, K.C., Nattee, C.: A comprehensive survey on on-line handwriting recognition technology and its real application to the nepalese natural handwriting (2009)
11. Tang, Y., Rose, R.: Rapid speaker adaptation using clustered maximum-likelihood linear basis with sparse training data. *IEEE Transactions on Audio, Speech, and Language Processing* 16(3), 607–616 (2008)
12. Baker, J.M., Deng, L., Glass, J., Khudanpur, S., Lee, C.-H., Morgan, N., OShaughnessy, D.: Research developments and directions in speech recognition and understanding, part 1. *IEEE Signal Processing Magazine* 26(3), 75–80 (2009)
13. Ohmura, R., Hashida, N., Imai, M.: Preliminary evaluation of personal adaptation techniques in accelerometer-based activity recognition. In: *Proc. 13th IEEE Int. Symposium on Wearable Computers: Late Breaking Results* (2009)
14. He, X., Zhao, Y.: Fast model selection based speaker adaptation for nonnative speech. *IEEE Trans. on Speech and Audio Processing* 11(4), 298–307 (2003)
15. Kunze, K., Lukowicz, P.: Using acceleration signatures from everyday activities for on-body device location. In: *2007 11th IEEE International Symposium on Wearable Computers*, September 2007, pp. 115–116 (2007)
16. Förster, K., Roggen, D., Tröster, G.: Unsupervised classifier self-calibration through repeated context occurrences: Is there robustness against sensor displacement to gain? In: *Proc. 13th IEEE Int. Symposium on Wearable Computers (ISWC)*, pp. 77–84 (2009)

17. Taylor, S.F., Stern, E.R., Gehring, W.J.: Neural systems for error monitoring: Recent findings and theoretical perspectives. *Neuroscientist* 13(2), 160–172 (2007)
18. Falkenstein, M., Hoormann, J., Christ, S., Hohnsbein, J.: ERP components on reaction errors and their functional significance: A tutorial. *Biol. Psychol.* 51(2-3), 87–107 (2000)
19. Ferrez, P.W., Millán, J.: Error-related EEG potentials generated during simulated brain-computer interaction. *IEEE Trans. Biomed. Eng.* 55, 923–929 (2008)
20. Schalk, G., Wolpaw, J.R., McFarland, D.J., Pfurtscheller, G.: EEG-based communication: Presence of an error potential. *Clin. Neurophysiol.* 111(12), 2138–2144 (2000)
21. Parra, L.C., Spence, C.D., Gerson, A.D., Sajda, P.: Response error correction—A demonstration of improved human-machine performance using real-time EEG monitoring. *IEEE Trans. Neural. Syst. Rehabil. Eng.* 11(2), 173–177 (2003)
22. Fatourechi, M., Bashashati, A., Ward, R.K., Birch, G.E.: EMG and EOG artifacts in brain computer interface systems: A survey. *Clin. Neurophysiol.* 118(3), 480–494 (2007)
23. Chavarriaga, R., Ferrez, P.W., Millán, J.: To Err Is Human: Learning from error potentials in brain-computer interfaces. In: International Conference on Cognitive Neurodynamics (2007)
24. Bollon, J.M., Chavarriaga, R., Millán, J., Bessière, P.: EEG error-related potentials detection with a Bayesian filter. In: 4th International IEEE EMBS Conference on Neural Engineering, Antalya Turkey (2009)
25. Gehring, W.J., Goss, B., Coles, M.G.H., Meyer, D.E., Donchin, E.A.: Neural system for error-detection and compensation. *Psychol. Sci.* 4, 385–390 (1993)
26. Schlögl, A., Keinrath, C., Zimmermann, D., Scherer, R., Leeb, R., Pfurtscheller, G.: A fully automated correction method of EOG artifacts in EEG recordings. *Clin. Neurophysiol.* 118(1), 98–104 (2007)
27. Liu, H., Setiono, R.: A probabilistic approach to feature selection - a filter solution, pp. 319–327. Morgan Kaufmann, San Francisco
28. García Lopez, F., García Torres, M., Melian Batista, B., Moreno Perez, J.A., Moreno-Vega, J.M.: Solving feature subset selection problem by a parallel scatter search. *European Journal of Operational Research* 169(2), 477–489 (2006)
29. John, G., Langley, P.: Estimating continuous distributions in Bayesian classifiers. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 338–345. Morgan Kaufmann, San Francisco (1995)
30. Castillo, E., Gutiérrez, J.M., Hadi, A.S.: Expert Systems and Probabilistic Network Models, Erste edn. Springer, New York (1996)
31. Aha, D.W., Kibler, D.: Instance-based learning algorithms. In: Machine Learning, pp. 37–66 (1991)
32. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, 1st edn. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, San Francisco (1999)
33. Tsymbal, A.: The problem of concept drift: Definitions and related work. Technical report, Department of Computer Science, Trinity College (2004)
34. Casson, A., Smith, S., Duncan, J., Rodriguez-Villegas, E.: Wearable EEG: what is it, why is it needed and what does it entail? In: Proc. IEEE Eng. Med. Biol. Soc., pp. 5867–5870 (2008)
35. Garipelli, G., Galán, F., Chavarriaga, R., Ferrez, P.W., Lew, E., Millán, J.: The use of Brain-Computer Interfacing for Ambient Intelligence. In: Intl. Workshop on Human Aspects in Ambient Intelligence (2007)

Author Index

- Abowd, Gregory D. 409
Acharya, Raviraja 94
Agarwal, Sharad 1
Amft, Oliver 319
Aoki, Paul 301

Bahl, Paramvir 1
Balazinska, Magdalena 57
Banerjee, Nilanjan 1
Baxi, Amit 94
Betz, Matthias 174
Bhattacharya, Sangeeta 94
Biasiucci, Andrea 427
Blunck, Henrik 38
Boll, Susanne 76
Bonner, Matthew N. 409
Borriello, Gaetano 57
Brown, Owain 210
Brudvik, Jeremy T. 409

Calabrese, Francesco 22
Campbell, Andrew T. 355
Chalmers, Matthew 210
Chandra, Ranveer 1
Chavarriaga, Ricardo 427
Cheng, Jingyuan 319
Christensen, Dan Lund 38
Cohn, Gabe 265
Conradi, Bettina 130
Corner, Mark 1
Corradi, Antonio 355

Darera, Vivek 94
del R. Millán, José 427
De Luca, Alexander 130
Deshpande, Piyush 94
Di Lorenzo, Giusy 22

Edwards, W. Keith 409
Englebienne, G. 283
Estrin, Deborah 138

Fodor, Kristof 355
Fogarty, James 57

Fürster, Kilian 427
Froehlich, Jon 265

Godsk, Torben 38
Grønbæk, Kaj 38
Gupta, Sidhant 228, 265

Hall, Malcolm 210
Healey, Jennifer 156
Herrmann, Klaus 373
Hodges, Mark R. 192
Hussmann, Heinrich 130

Ishiguro, Katsuhiko 246

Kamei, Koji 246
Kientz, Julie A. 228
Kirsch, Ned L. 192
Kishino, Yasue 246
Kjærgaard, Mikkel Baun 38
Kodalapura, Nagaraju 94
Kröse, B.J.A. 283
Kumar, Neil 301

Larson, Eric 265
Liu, Liang 22
Lukowicz, Paul 319

Maekawa, Takuya 246
Mageshkumar, Vincent 94
McMillan, Donald 210
Morris, Margaret 156
Morrison, Alistair 210
Musolesi, Mirco 355

Nachman, Lama 94, 156
Newman, Mark W. 192

Okadome, Takeshi 246
Oki, Maho 112

Patel, Shwetak N. 228, 265
Pereira, Francisco C. 22
Pielot, Martin 76

- Piraccini, Mattia 355
Pollack, Martha E. 192
Poupyrev, Ivan 391
- Rath, Satish 94
Ratti, Carlo 22
Reddy, Sasank 138
Roggen, Daniel 427
Rothermel, Kurt 373
- Sakurai, Yasushi 246
Schuhmann, Stephan 373
Seifert, Julian 130
Shahabdeen, Junaith 94, 156
Srinivasan, Vijay 337
Srivastava, Mani 138
- Stankovic, John 337
Subramanian, Sushmita 156, 301
- Toftkjær, Thomas 38
Tröster, Gerhard 427
Tsukada, Koji 112
- van Kasteren, T.L.M. 283
- Welbourne, Evan 57
Whitehouse, Kamin 337
Willett, Wesley 301
Willis, Karl D.D. 391
Wolman, Alec 1
Woodruff, Allison 301
- Yanagisawa, Yutaka 246